Cross-Document Cross-Lingual NLI via RST-Enhanced Graph Fusion and Interpretability Prediction

Anonymous ACL submission

Abstract

Natural Language Inference (NLI) is a fundamental task in natural language processing. While NLI has developed many sub-directions such as sentence-level NLI, document-level NLI and cross-lingual NLI, Cross-Document Cross-Lingual NLI (CDCL-NLI) remains largely unexplored. In this paper, we propose a novel paradigm: CDCL-NLI, which extends traditional NLI capabilities to multi-document, multilingual scenarios. To support this task, we construct a high-quality CDCL-NLI dataset including 25,410 instances and spanning 26 languages. To address the limitations of previous methods on CDCL-NLI task, we further propose an innovative method that integrates RST-enhanced graph fusion with interpretability-aware prediction. Our approach leverages RST (Rhetorical Structure Theory) within heterogeneous graph neural networks for cross-document context modeling, and employs a structure-aware semantic alignment based on lexical chains for cross-lingual understanding. For NLI interpretability, we develop an EDU (Elementary Discourse Unit)-level attribution framework that produces extractive explanations. Extensive experiments demonstrate our approach's superior performance, achieving significant improvements over both conventional NLI models as well as large language models. Our work sheds light on the study of NLI and will bring research interest on cross-document cross-lingual context understanding, hallucination elimination and interpretability inference. Our code and datasets are available at CDCL-NLI-link for peer review.

1 Introduction

Natural Language Inference (NLI) is a fundamental task in natural language processing, aiming to determine the logical relationship between the given premise and hypothesis pair (Dagan et al., 2005; MacCartney and Manning, 2009). While traditional NLI tasks primarily deal with single-language, short-text validations (Rodrigo et al., 2007), document-level NLI (Yin et al., 2021) expands the scope of NLI to longer contexts.

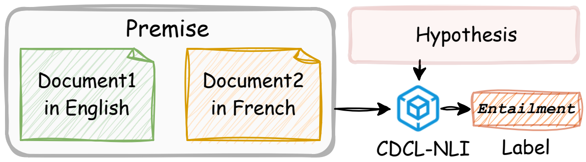


Figure 1: A CDCL-NLI example. Premise in English and French. The Entailment label requires combining information from both documents in premise.

Paradigm	Premise	Hypothesis	Language
Sentence-NLI	Sentence	Sentence	Mono/Multi
Document-NLI	Doc	Sent/Doc	Mono
CDCL-NLI	Multi Doc	Sentence	Multi

Table 1: Comparison of different NLI paradigms.

Table 1 compares different NLI paradigms systematically, highlighting the progressive evolution of NLI tasks. Sentence-NLI involves low-complexity reasoning on short sentence pairs, evolves from single-language approaches (Bowman et al., 2015; Herlihy and Rudinger, 2021) to multilingual settings (Conneau et al., 2018; Heredia et al., 2024), and is mainly used for fact verification (Wadden et al., 2020; Klemen et al., 2024). Document-level NLI extends NLI to reasoning over full-length documents within a single language (Wang et al., 2019; Yin et al., 2021), focusing on content comprehension (Yang et al., 2024).

However, the increasing globalization of information flow requires even more sophisticated inference capabilities across both language and document boundaries. In this paper, we introduce **C**ross-**D**ocument **C**ross-**L**ingual **N**atural **L**anguage **I**nference (**CDCL-NLI**), a novel paradigm extending traditional NLI to multi-document and multilingual settings. Figure 1 illustrates that CDCL-NLI jointly reasons over premise documents in English and French to verify the hypothesis. The correct

Entailment prediction relies on integrating complementary information from both documents.

While CDCL-NLI addresses a real-world task with broad applications, it faces key challenges: **1) Lack of existing datasets**, which necessitates the construction of new resources to support research. **2) Multilingual Semantic Alignment**, requiring resolution of grammatical and conceptual differences across languages while preserving semantic consistency (Conneau et al., 2020). **3) Cross-Document Structure Alignment**, essential for capturing structural correspondences and implicit logical relations between documents of varying complexity (Wang et al., 2021); and **4) Interpretability**, demanding transparent reasoning processes and verifiable confidence in inference outcomes (Bereska and Gavves, 2024).

To address the first challenge, we curated a **CDCL-NLI dataset** through collecting diverse premise documents from GlobeSumm (Ye et al., 2024), generating hypotheses with GPT-4o (OpenAI, 2024) using customized prompts to ensure label diversity and balance and manually reviewing hypotheses and annotated explanations. The dataset contains 25,410 samples spanning 26 languages and 370 events.

To address the rest challenges, we proposed a novel method that comprises three key components. **1) Graph Construction Module:** This component promotes semantic alignment by fusing graphs based on lexical chains, effectively linking semantically related concepts across documents. **2) Graph Representation Module:** Utilizing an RST-enhanced Relation-aware Graph Attention Network (RGAT) (Mann and Thompson, 1988; Busbridge et al., 2019), this module supports structure alignment by capturing hierarchical discourse structures and cross-document dependencies through multi-head attention mechanisms. **3) Interpretability Attribution Module:** Leveraging Elementary Discourse Units (EDUs) (Mann and Thompson, 1988), this module generates extractive explanations that significantly enhance model interpretability and provide transparent insights into its decision-making process.

Extensive experiments on the CDCL-NLI and DocNLI datasets demonstrate that our method outperforms conventional NLI approaches and three state-of-the-art large language models, surpassing the strongest baseline by 3.5% on our dataset. In the end, we highlight our main contributions as follows:

- We propose CDCL-NLI as a new task and construct a corresponding dataset covering 26 languages with 25,410 high-quality manually-annotated instances.
- We propose a novel method, which leverages RST-enhanced graph fusion to align semantic concepts and discourse structures, and improves interpretability by generating extractive explanations based on EDUs.
- We conduct extensive experiments, outperforming all the baselines by at least 3.5%.

2 Related Work

2.1 Sentence-level NLI

Monolingual Methods. Sentence-level NLI benchmarks like SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) have driven model evolution from ESIM (Chen et al., 2017) to transformer architectures (Devlin et al., 2018; Liu et al., 2019) and recent LLMs (OpenAI, 2023).

Cross-lingual Methods. Cross-lingual NLI relies on datasets like XNLI (Conneau et al., 2018) (15 languages) and XNLIeu (Heredia et al., 2024) (European languages). Multilingual models such as XLM-R (Conneau et al., 2020) and XLM-E (Chi et al., 2022) enable zero-shot transfer, while alignment methods like SoftMV (Hu et al., 2023) and prompt-based MPT (Qiu et al., 2024) improve cross-lingual semantic understanding.

Interpretability Mechanisms. Interpretability uses feature attribution methods like Integrated Gradients (Sundararajan et al., 2017) and (Huang et al., 2024) to highlight decision-driving features. Datasets such as e-SNLI (Camburu et al., 2018) provide human explanations, supporting explicit reasoning and interpretability benchmarks.

2.2 Document-level NLI

Datasets and Benchmarks. Document-level NLI benefits from datasets like DocNLI (Yin et al., 2021) with over one million instances. Domain-specific datasets such as ContractNLI (Koreeda and Manning, 2021) focus on the challenges posed by long documents and specialized text genres.

Inference Methods. Recent approaches emphasize discourse structure and long-range dependencies. R2F (Wang et al., 2022) introduces explicit reasoning extraction, and DocInfer (Mathur et al., 2022) uses hierarchical encoding to model docu-

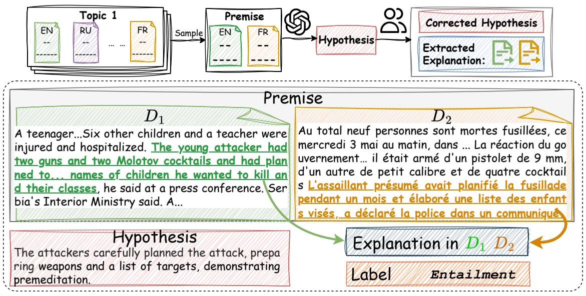


Figure 2: Overview of the CDCL-NLI dataset construction process and a data example. Premise contains D_1 and D_2. Explanation is extracted from premise to enhance interpretability.

ment structure, highlighting the need to capture document-level semantics.

Interpretability Mechanisms. Interpretability research focuses on evidence extraction and explanation generation. Systems like EvidenceNet (Chen et al., 2022) and R2F (Wang et al., 2022) automatically identify evidence to enhance reasoning transparency. LLM-based approaches like Chain-of-Thought (Wei et al., 2022) and Rethinking(Singh et al., 2024) further enable self-explanatory reasoning capabilities.

Although prior studies have advanced sentence-level and document-level NLI, challenges in cross-document and cross-lingual NLI remain largely unaddressed. Our work fills this gap by introducing the CDCL-NLI dataset and proposing an interpretable RST-enhanced graph fusion method.

3 CDCL-NLI Task Formulation and Dataset Construction

As shown in Figure 2, our CDCL-NLI dataset is constructed through a systematic pipeline involving stratified random sampling of premise documents across all topics, LLM-generated hypotheses, and human verification to ensure data quality. In the dashed box, the figure shows a CDCL-NLI instance with a premise of two documents in different languages, an English hypothesis, a label, and EDU-based explanations for interpretability.

3.1 Task Formulation

Similar to the traditional NLI task, the goal of CDCL-NLI is to determine the inference label:

$$Label \in \{\textit{"Entailment"}, \textit{"Neutral"}, \textit{"Contradiction"}\},$$

between a given premise P and hypothesis H. Specifically, the premise P consists of two documents D_1 and D_2, written in different languages but discussing the same topic. The hypothesis H is a sentence-level statement. The task requires reasoning over the combined information from P with H to determine their entailment relationship, involving both cross-document and cross-lingual premise integration.

3.2 Premise Data Collection

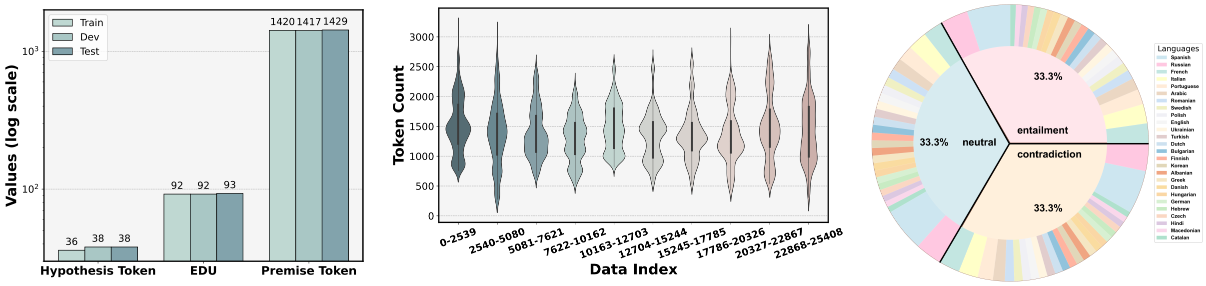
We collect our premises from GlobeSumm (Ye et al., 2024), a multi-document cross-lingual summarization dataset covering 370 topics across 26 languages. In GlobeSumm, documents for each topic span diverse media outlets, publication times, and languages, providing a rich foundation for cross-document and cross-lingual inference tasks. We curated CDCL-NLI dataset by stratified randomly selecting documents for each topic to form premise pairs. To enhance cross-lingual coverage, we strategically expanded our document collection through translation.After rigorous quality filtering, our final dataset consists of high-quality inference instances covering 26 language combinations. Detailed premise establishment criteria and quality filtering standards are provided in Appendix A.1.

3.3 Hypothesis Generation and Label Specification

For each pair, we generate hypotheses across three NLI categories. Initial hypotheses are generated by GPT-4o (OpenAI, 2024) following specific guidelines to ensure balanced label distribution and sufficient reasoning depth. Entailment hypotheses require joint or consistent support from the premise documents. Neutral hypotheses are plausible but neither supported nor contradicted. Contradiction hypotheses explicitly conflict, focusing on cross-document inconsistencies. To reduce hallucination, GPT-4o first generates explanations before finalizing hypotheses. Detailed prompts and protocols are included in Appendix A.2.

3.4 Manual Annotation and Quality Control

Our annotation involved two phases: hypothesis verification and EDU-based explanation (Figure 2). Three graduate students independently labeled premise-hypothesis pairs, achieving strong inter-annotator agreement (Cohen's κ: 0.71–0.82 across classes). For explanations, annotators selected minimal EDU sets supporting their decisions, with high agreement (Jaccard: 0.91; span overlap: 0.94; conclusion: 1.00). All annotations were reconciled through discussions to ensure quality (see Appendix A.3).



(a) Log-scaled token count and EDU numbers across different data splits.

(b) Premise token length distributions across different data index groups.

(c) Label distribution with language composition.

Figure 3: Statistic visualization of token length, EDU numbers, label distribution and language composition.

Dataset	CD	CL	Interp.	Avg.Tks	Labels
MultiNLI	×	×	×	33.7	3
XNLI	×	✓	×	50	3
e-SNLI	×	×	✓	45.1	3
DocNLI	✓	×	×	412	2
CDCL-NLI	✓	✓	✓	1,456	3

Table 2: Characteristics of NLI datasets showing cross-document (CD), cross-lingual (CL), and interpretability (Interp.) capabilities, along with average tokens per instance (Avg.Tks) and number of label classes.

3.5 Dataset Statistics

We summarize the key characteristics of different NLI datasets in Table 2, which shows substantial variations in their cross-document and cross-lingual capabilities. Our CDCL-NLI dataset consists of 25,410 cross-document, cross-lingual NLI instances spanning 26 languages and 370 events. We partitioned the dataset by event topics, yielding 22,200/1,605/1,605 train/dev/test instances with mutually exclusive event distributions. Figure 3a shows similar data characteristics across training, validation, and test sets; Figure 3b depicts token count variations across consecutive segments; and Figure 3c illustrates balanced label distributions (33.3% each) with roughly uniform language distribution within each label. We provide more information about our dataset in Appendix A.4.

4 Our Method: RST-enhanced Graph Fusion with EDU Level Interpretability

Our approach offers a robust solution for cross-document and cross-lingual NLI by leveraging RST-enhanced graph fusion and explanation prediction. As illustrated in Figure 4, the framework comprises three main components: RST graph construction and fusion module, graph representation generation module and interpretability and classification module.

4.1 RST Graph Construction and Fusion

RST Information Extraction. We employ DM-RST (Liu et al., 2021), a top-down multilingual document-level rhetorical structure parsing framework, to extract RST information from the premise documents. As shown in Figure 5, DM-RST generates two key features for document D: 1) EDU boundary indices and 2) RST tree parsing outputs. By processing these features, we get $D = \{EDU_1, EDU_2, ..., EDU_n\}$ and rhetorical structure tree $\mathcal{T}$. EDU_i represents the i-th EDU's textual content. $\mathcal{T}$ is formally defined as:

$$\mathcal{T} = \left\{ \begin{matrix} (EDU_{[s \to t]}, EDU_{[t+1 \to u]}, r_{st}, r_{tu}) \mid \\ s, t, u \in [1, n],\ s \le t < u,\ r_{st}, r_{tu} \in \mathcal{R} \end{matrix} \right\},$$

where $EDU_{[s \to t]}$ denotes an EDU group that forms either a leaf node (when $s = t$) or a branch node (when $s < t$), and r_{st} represents the rhetorical relation. This tree structure captures both local EDU relationships and global discourse organization.

Embedding Model. To handle inconsistent cross-lingual encoding from premise documents in different languages, we use XLM-RoBERTa-Large (Conneau et al., 2020) as the base encoder, which supports over 100 languages and excels at multilingual semantic representation. For each EDU_i in the RST structure, its initial vector is $\boldsymbol{h}_{EDU_i} = \phi(EDU_i) \in \mathbb{R}^d$, where ϕ denotes XLM-RoBERTa-Large and $d = 1024$. The hypothesis vector $\boldsymbol{h}_{\text{hypo}}$ is computed similarly.

Single Graph Construction. Based on the RST tree $\mathcal{T}$, we construct graphs G_{D_1} and G_{D_2} for each document D_1 and D_2 respectively as shown in Figure 4. For graph $G(V, E, R)$, we define:

- **Node Set** $\mathcal{V} = \{v_i \mid \text{EDU}_{[s \to t]} \in \mathcal{T}\}$, where each v_i has features: Text_{v_i}, ϕ_{v_i}, and Type_{v_i}.
- **Edge Set** $\mathcal{E} = \{(v_i, v_j) \mid v_i \neq v_j, (v_i, v_j, r) \in \mathcal{T}\}$, representing bidirectional edges.

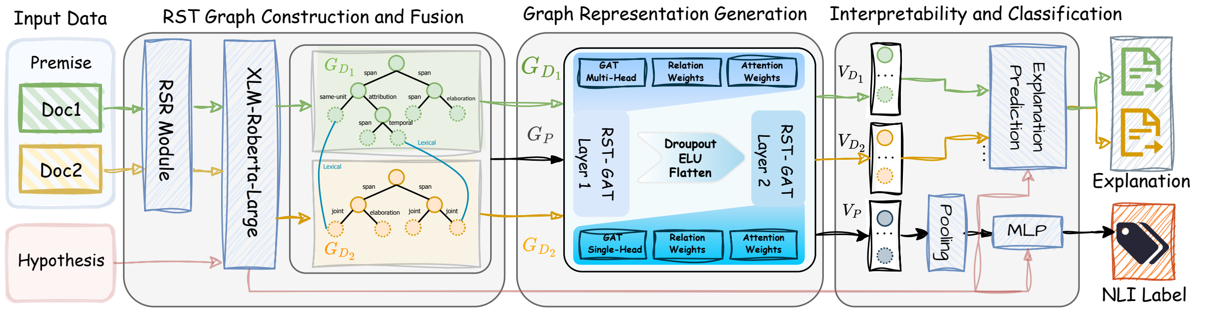


Figure 4: Our CDCL-NLI framework processes premise (D_1, D_2) and hypothesis through: 1) RST Graph Construction, merging G_{D_1} and G_{D_2} into G_P; 2) Graph Representation via *RST-GAT* layers; 3) Interpretability and Classification, extracting node-level explanations while using $\boldsymbol{h}_{G_p}$ and $\boldsymbol{h}_{hypo}$ for final NLI label prediction.

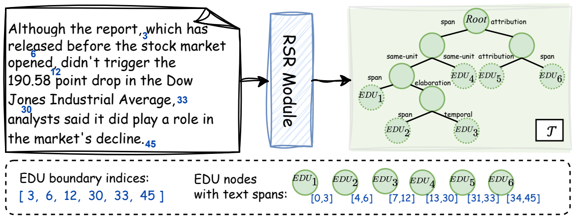


Figure 5: RST graph construction. The RST module first segments text into EDUs(EDU_1-EDU_6), with boundaries in blue, and then organizes an RST tree $\mathcal{T}$ showing discourse relations.

- **Relation Set** $\mathcal{R}$ is from rhetorical relations in $\mathcal{T}$. For detailed relations and definitions of node features, please refer to the Appendix B.1, B.2.

Graph Fusion. After obtaining heterogeneous graphs $G_{D_1}(V_{D_1}, E_{D_1}, R)$ and $G_{D_2}(V_{D_2}, E_{D_2}, R)$ for the premise, we then merge them via lexical chains to enhance cross-document reasoning by:

- **Node Feature Fusion**: $V_P = V_{D_1} \cup V_{D_2}$, retaining all nodes and features.
- **Cross-document Edge**: Add bidirectional lexical edges between $v_i \in V_{D_1}$ and $v_j \in V_{D_2}$ if $\text{CosineSim}(v_i, v_j) > \delta$, and obtain E_P.[1]
- **Adding Edge Types**: Extend R with a new "Lexical" relation R' to support lexical alignment.

The merged graph $G_P(V_P, E_P, R')$ preserves individual features while aligning semantics across documents, effectively supporting CDCL-NLI.

4.2 Graph Representation Generation

Node-level Representation. As shown in Figure 4, there are two layers of *RST-GAT* to process nodes' features. *RST-GAT* builds upon the Relation-aware Graph Attention Network (RGAT) (Busbridge et al., 2019), which extends Graph Attention Network (GAT) (Velickovic et al., 2018) to handle relation-specific edge types in graphs.

Taking a graph $G(V, E, R)$ as an example, the initial node embeddings $\mathbf{h}_V^0$ are obtained as described in Section 4.1. Node representations are then updated through two layers of relation-aware multi-head attention as follows:

$$\mathbf{h}_{v_i}^{(l)} = \frac{1}{|R|}\sum_{r\in R}\alpha_r \cdot \frac{1}{K}\sum_{k=1}^{K}\sum_{v_j\in\mathcal{N}_r(v_i)} \beta_{ij,k}^{r,(l)}\mathbf{W}_{r,k}\mathbf{h}_{v_j}^{(l-1)} \quad (1)$$

where $l = 1, 2$. Here, α_r denotes the softmax-normalized weight of relation r, capturing the relative importance among relations, while $\beta_{ij,k}^{r,(l)}$ represents the attention coefficient over neighboring nodes, indexed by node pairs (v_i, v_j), attention head k, relation r, and layer l. After two layers of message passing, the resulting node embeddings are denoted as $\mathbf{h}_V = \{\mathbf{h}_{v_i}^{(2)}\}$. The same update procedure is applied independently to G_{D_1}, G_{D_2}, and G_P, producing embeddings $\mathbf{h}_{V_{D_1}}$, $\mathbf{h}_{V_{D_2}}$, and $\mathbf{h}_{V_P}$, respectively. Detailed formulations of the attention weights and parameter configurations are provided in Appendix B.4.

Graph-level Representation. The global representation($\boldsymbol{h}_{G_P}$) of the merged graph G_P is obtained by averaging node features after two *RST-GAT* layers. This pooling captures discourse-level semantics while preserving local rhetorical relations, enabling effective classification.

Classification Loss. Given the concatenated graph representation $\boldsymbol{h}_{G_p}$ and hypothesis features $\boldsymbol{h}_{hypo}$, the classification loss is computed using the standard cross-entropy (CE) formulation:

$$\mathcal{L}_{\text{cls}} = \text{CE}(\mathbf{y}, \text{Softmax}(\text{MLP}(\boldsymbol{h}_{G_p} \oplus \boldsymbol{h}_{hypo})) \in \mathbb{R}^3), \quad (2)$$

where $\mathbf{y}$ denotes the ground-truth label and $\mathbf{p}$ denotes the predicted probability distribution.

[1] Threshold δ is chosen empirically; see Appendix B.3 for detailed justification.

Enhanced Triplet Loss. Triplet loss (Weinberger and Saul, 2006; Schroff et al., 2015) is a metric learning method that encourages the anchor-positive distance to be smaller than the anchor-negative distance. Leveraging the structure of our CDCL-NLI dataset, where each premise aligns with three hypotheses (entailment, neutral, contradiction), we propose a neutral-constrained triplet loss:

$$\begin{aligned}\mathcal{L}_{\text{triplet}} &= \max(0, d(a,p) - d(a,n) + \sigma) \\ &\quad + \max(0, d(a,\text{neu}) - d(a,n) + \theta),\end{aligned} \tag{3}$$

where $d(x, y)$ is the Euclidean distance, and a, p, neu, n denote the premise paired with entailment, neutral, and contradiction hypotheses, respectively. Margins σ and θ enforce the semantic order: entailment < neutral < contradiction.

4.3 EDU-level Explanation Prediction

For interpretability, we propose an attention-based method to extract explanation nodes.

Node Importance. Using multi-head attention weights from the first *RST-GAT* layer, the importance score I_i of node v_i in G_{D_1}, G_{D_2} is

$$I_i = \frac{1}{K}\sum_{k=1}^{K}\sum_{r\in R}\sum_{v_j\in\mathcal{N}_r^{\text{in}}(v_i)}\beta_{ji,k}^{r,(1)}. \tag{4}$$

Let $\boldsymbol{H} = [\boldsymbol{h}_{v_0}; \ldots; \boldsymbol{h}_{v_n}]$ be node features and $\boldsymbol{I} = [I_0, \ldots, I_n]^\top$ importance scores. Weighted features are $\boldsymbol{H}' = \boldsymbol{I} \odot \boldsymbol{H}$, where $\odot$ denotes element-wise product with broadcasting.

Hypothesis-aware Interaction. Given hypothesis embedding $\boldsymbol{h}_{hypo} \in \mathbb{R}^{d^{\text{out}}}$, attention over weighted features $\boldsymbol{H}' \in \mathbb{R}^{n\times d^{\text{out}}}$ produces interaction features:

$$\boldsymbol{O} = \text{Attention}\left(\frac{\boldsymbol{h}_{hypo}\boldsymbol{H}'^{\top}}{\sqrt{d^{\text{out}}}}\right)\boldsymbol{H}'. \tag{5}$$

Feature Fusion and Classification. The model is optimized by Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{\text{exp}} = \frac{1}{N}\sum_{i=1}^{N}\text{BCE}\big(y_i, \text{Sigmoid}(\text{MLP}([\boldsymbol{h}_i' \oplus \boldsymbol{o}_i]))\big) \tag{6}$$

where $y_i \in \{0, 1\}$ is ground truth label of node i, $\boldsymbol{h}_i'$ and $\boldsymbol{o}_i$ are the weighted and interaction features for node i respectively.

The total loss combines all components:

$$\mathcal{L}_{\text{total}} = \gamma\mathcal{L}_{\text{exp}} + \lambda(\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{triplet}}), \tag{7}$$

where γ and λ are balancing hyperparameters set as 0.2 and 0.8 respectively through grid search on the validation set.

5 Experiments

5.1 Experiment Settings

Metrics. Model evaluation considers classification and explanation quality. For classification on DocNLI (imbalanced), we report **Micro F1** and **Weighted F1**. On CDCL-NLI dataset, we use **Macro Precision**, **Macro Recall**, and **Macro F1** for balanced class performance. Explanation quality is assessed using **BLEU** (1-4), **ROUGE-1/2/L**, and **METEOR**.

Baselines.

- **Conventional NLI Models**: We compare two well-established models, both trained on our dataset: **DocNLI** (Yin et al., 2021), a document-level NLI model tailored for long texts, and **R2F** (Wang et al., 2022), a retrieval-based framework for document-level NLI. All conventional baselines and our proposed method are built upon the same underlying pretrained language model to ensure fair comparison. Training details are provided in Appendix C.1.
- **Large Language Models**: We evaluate three LLMs: **Llama3-8B-Instruct** (Meta AI, 2024), **Qwen-3-8B** (Qwen, 2025) and **GPT-4o** (OpenAI, 2024), where the LLaMA and Qwen model is further fine-tuned with LoRA adapters. All models are tested in a few-shot setting, with fine-tuning configurations in Appendix C.2.

5.2 Experiment Results and Analysis

Main Results and Ablation Study. Table 3 presents a performance comparison of our proposed method against several competitive baselines on two test sets. TestSet1 is a cross-lingual test set (the original test set of the CDCL-NLI dataset). TestSet2 is an English-translated version of TestSet1, designed to evaluate model robustness in a cross-document scenario without language barriers, and to quantify the performance degradation caused by cross-lingual factors. This dual evaluation framework enables a clearer analysis of the impact of language variation on NLI performance.[2]

Our model consistently achieves the best results on both test sets, with macro F1 scores of **68.95%** on the cross-lingual set and **70.68%** on the English-translated set, surpassing strong baselines such as DocNLI and R2F by notable margins. The generally higher scores on the English test set highlight the relative ease of reasoning within

[2]Unless noted, all reported test results refer to TestSet1.

Model Type	Model	TestSet1:Cross-Lingual			TestSet2:English			Trained
		Precision	Recall	F1 Macro	Precision	Recall	F1 Macro	
Conventional Model	Hypothesis-only	35.78	36.02	35.84	35.89	35.97	36.12	✓
	DocNLI	64.75	64.30	64.46	69.29	68.39	68.70	✓
	R2F	65.04	65.42	65.42	67.18	68.47	67.13	✓
Large Language Model	Llama-3-8B	45.94	52.62	48.07	51.69	57.98	53.03	✓
	GPT-4o	52.50	56.30	54.00	62.50	65.00	64.50	×
	Qwen3-8B	60.34	56.29	59.86	71.71	67.62	67.34	✓
CDCL-NLI Model	Ours	**71.09**	**70.84**	**68.95**	**72.65**	**72.46**	**70.68**	✓
	- Exp	65.99	67.29	65.86	69.01	69.97	68.79	✓
	- Graph	53.07	57.38	51.37	68.64	64.55	61.71	✓
	- Exp & Graph	49.15	52.71	48.70	49.15	52.71	53.29	✓

Table 3: NLI model performance on cross-lingual (TestSet1) and English (TestSet2) sets. Our full model achieves the highest F1 scores, showing clear gains from explanation and graph components. Large language models perform well but are generally outperformed. ✓ indicates training on target data; × means no training. Explanation - Exp.

a single, well-resourced language, in contrast to the added challenges of cross-lingual understanding, which requires effective language transfer and alignment. The hypothesis-only baseline, which trains solely on the hypothesis, attains near-random performance (36% F1), indicating minimal dataset artifacts in the hypothesis statements.

Among the large language models evaluated in the few-shot setting, Qwen3-8B achieves the best performance, with F1 scores of 59.86% on the cross-lingual set and 67.34% on the English set, outperforming both GPT-4o and Llama3-8B. Nevertheless, our approach surpasses Qwen3-8B by 9.09% on the cross-lingual set and 3.34% on the English set, highlighting the effectiveness of our method. Detailed prompts and zero-shot results and reported in Appendix D.1, Appendix D.2.

The ablation study highlights the importance of each component: removing the explanation module (- Exp) results in a moderate performance drop of 1.89% on both cross-lingual and English test sets; removing the graph module (- Graph) causes a more pronounced decline of 17.58% and 8.97%, respectively. When both components are removed (- Exp & Graph), performance sharply decreases on both test sets, demonstrating that these modules jointly contribute to the model's robustness under different language conditions.

Single-Document vs Cross-Document. To validate the cross-document nature of our dataset, we compare the performance of models using only a single document ($D1$ or $D2$) against those using the $D1 + D2$, as illustrated in Figure 6. The substantial performance gap—at least a 7% F1 improvement—demonstrates that effective inference requires integrating information from both documents. Additionally, the similar F1 scores for $Document_1$ (63.2%) and $Document_2$ (62.8%) indicate that both documents provide equally important information, underscoring the necessity of synthesizing evidence from both sources rather than relying on either alone. Additional results are presented in Appendix D.3.

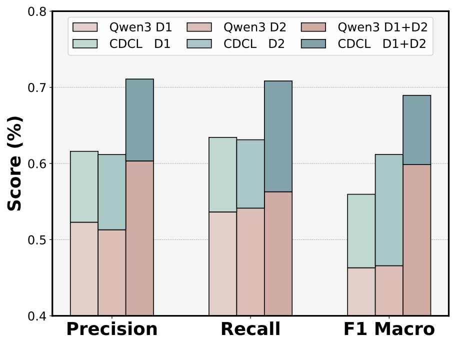


Figure 6: NLI performance using single documents ($D1$, $D2$) versus combined ($D1 + D2$). The F1 gain confirms the need for cross-document reasoning, with both documents contributing similarly.

Cross-Lingual Generalization. To further assess the robustness and generalization of our approach, we conduct cross-lingual transfer experiments in a challenging scenario where the training and testing languages are distinct. Specifically, we select five typologically and geographically diverse languages—Spanish, Russian, French, Italian, and English—to ensure comprehensive coverage and to reflect real-world multilingual settings. For each source language, we translate the data into all target languages, resulting in 20 transfer directions. Models are trained on one language and evaluated on a different target language, with no overlap between training and test languages. As shown in Table 4, our method consistently outperforms the

F1 Scores on Target Language (Ours vs. R2F)			
ES→RU	ES→FR	ES→IT	ES→EN
55.53/25.03	**58.28**/27.31	**54.68**/29.31	**57.94**/34.21
RU→ES	RU→FR	RU→IT	RU→EN
52.83/46.26	**46.67**/35.50	**50.89**/39.77	**49.67**/47.78
FR→ES	FR→RU	FR→IT	FR→EN
50.31/43.25	**56.6**/22.24	**58.65**/39.32	**49.67**/47.22
IT→ES	IT→RU	IT→FR	IT→EN
53.72/36.01	**57.19**/36.21	**53.17**/37.22	**56.67**/47.21
EN→ES	EN→RU	EN→FR	EN→IT
60.31/49.94	**51.27**/32.46	**60.28**/30.80	**55.11**/38.33

Table 4: Cross-lingual performances (macro F1 scores) of our method and R2F. Source languages are colored. Spanish (ES), Russian (RU), French (FR), Italian (IT) and English (EN). Our method demonstrates superior generalization across languages compared to baselines.

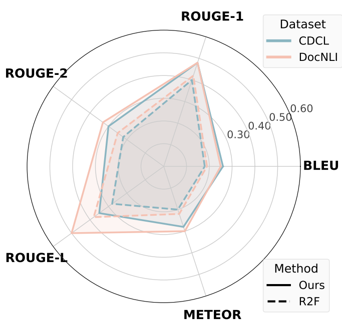


Figure 7: Explainability comparison between our method and R2F on CDCL-NLI and DocNLI datasets using BLEU, ROUGE (1/2/L), and METEOR metrics. Our method consistently outperforms R2F across all metrics and datasets.

R2F baseline across most transfer directions, often by substantial margins. R2F is chosen as it improves upon DocNLI for cross-document reasoning. These results demonstrate the effectiveness of our approach in synthesizing information from cross-lingual document pairs and its strong transferability to diverse language pairs, validating the design of our experimental setup and the broad applicability of our method in multilingual cross-document NLI tasks.

Interpretability Study. To evaluate our method's effectiveness, we compared it against the R2F baseline using five standard metrics (ROUGE-1/2/L, BLEU, METEOR) on both CDCL and DocNLI datasets. As shown in Figure 7, our method (solid line) consistently outperforms r2f (dashed line) across all metrics on both datasets. The improvements are particularly pronounced in ROUGE-L, where our method achieves 0.34 versus 0.30 on CDCL-NLI and 0.50 versus 0.37 on DocNLI, demonstrating enhanced capability in preserving structural coherence. It is worth noting that the interpretability data for DocNLI was provided by R2F.

Method	Dev		Test	
	W. F1	Mi. F1	W. F1	Mi. F1
DocNLI	88.05	86.25*	87.09	85.06*
R2F	90.18*	**89.15**	89.16*	87.86
Ours	**91.58**	88.61	**90.30**	**88.47**

Table 5: Performance comparison on the document-level DocNLI. Results marked with * are from our reproduction. Weighted F1 -W. F1, Micro F1 - Mi. F1

Comparison on DocNLI Dataset. We evaluate the generalization of our method on the DocNLI dataset using weighted and micro F1 metrics. As shown in Table 5, our approach achieves state-of-the-art weighted F1, outperforming both the DocNLI baseline and R2F, but slightly underperforms R2F on micro F1. This is mainly due to class imbalance between training and evaluation sets, and R2F's advantage on the simpler reasoning tasks common in DocNLI, while our method is optimized for more complex reasoning. These results suggest that balanced sampling or improved adaptability could further boost performance.

6 Conclusion

This work systematically investigates CDCL-NLI, addressing key challenges in cross-document reasoning and multilingual understanding. We introduce a novel CDCL-NLI dataset spanning 26 languages and comprising 25,410 meticulously annotated instances. And we propose an RST-enhanced graph fusion mechanism with explanation prediction. Through extensive experiments and analyses, we demonstrate that our method effectively captures both structural and semantic information across documents and languages. Specifically, the RST-enhanced graph fusion mechanism and explanation prediction component not only improve model interpretability but also enhance performance, as validated by our ablation studies. Our empirical findings provide several key insights: 1) integrating rhetorical structure significantly improves the graph model's ability to capture document-level discourse information; 2) reasoning cross-document is necessary and our method has strong cross-lingual reasoning capability; and 3) the EDU-level attribution method has a beneficial effect on classification and could generate explanations aligned with human reasoning.

Limitations

Our current framework is constrained to reasoning between pairs of documents, while real-world scenarios often involve multiple documents across diverse topics. This limitation points to valuable directions for future research in multi-document multi-lingual inference.

Ethics Statement

All data in our proposed dataset are collected from publicly available sources with respect for privacy and copyright. We have removed any personally identifiable information during preprocessing. The dataset is intended for research purposes only, and we advise users to be aware of potential biases present in the original data.

References

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. The Association for Computational Linguistics.

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y. Hammerla. 2019. Relational graph attention networks. *CoRR*, abs/1904.05811.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. The Association for Computational Linguistics.

Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Fei Li, Meihuizi Jia, and Jiaqi Li. 2022. Evidencenet: Evidence fusion network for fact verification. In *Proceedings of the ACM Web Conference 2022*, pages 2636–2645. ACM.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6170–6182. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. The Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. Xnlieu: a dataset for cross-lingual nli in basque. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4177–4188. The Association for Computational Linguistics.

Christine Herlihy and Rachel Rudinger. 2021. Mednli is not immune: Natural language inference artifacts in the clinical domain. *arXiv preprint arXiv:2106.01491*.

Xuming Hu, Aiwei Liu, Yawen Yang, Fukun Ma, S Yu Philip, Lijie Wen, and 1 others. 2023. Enhancing cross-lingual natural language inference by soft prompting with multilingual verbalizer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1361–1374. The Association for Computational Linguistics.

Guangming Huang, Yingya Li, Shoaib Jameel, Yunfei Long, and Giorgos Papanastasiou. 2024. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Computational and Structural Biotechnology Journal*.

Matej Klemen, Ales Zagar, Jaka Cibej, and Marko Robnik-Sikonja. 2024. SI-NLI: A slovene natural language inference dataset and its evaluation. In *Proceedings of the 2024 Joint International Conference* *on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 14859–14870. ELRA and ICCL.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919. The Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zhengyuan Liu, Ke Shi, and Nancy F. Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. *CoRR*, abs/2110.04518.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Puneet Mathur, Gautam Kunapuli, Riyaz Bhat, Manish Shrivastava, Dinesh Manocha, and Maneesh Singh. 2022. Docinfer: Document-level natural language inference using optimal evidence selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 809–824. The Association for Computational Linguistics.

Meta AI. 2024. Llama 3: Open Foundation and Fine-Tuned Chat Models. https://ai.meta.com/llama/. Accessed: 2024-03.

OpenAI. 2023. Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf.

OpenAI. 2024. GPT-4 Technical Report. https://openai.com/gpt-4. Accessed: 2024-03.

Xiaoyu Qiu, Yuechen Wang, Jiaxin Shi, Wengang Zhou, and Houqiang Li. 2024. Cross-lingual transfer for natural language inference via multilingual prompt translator. *arXiv preprint arXiv:2403.12407*, pages 1–6.

Team Qwen. 2025. Qwen3.

Álvaro Rodrigo, Anselmo Peñas, Jesús Herrera, and Felisa Verdejo. 2007. Experiments of UNED at the third recognising textual entailment challenge. In *Proceedings of the ACL-PASCAL@ACL 2007 Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic, June 28-29, 2007*, pages 89–94. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, PMLR.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021. Structure-augmented text representation learning for efficient knowledge graph completion. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1737–1748. ACM / IW3C2.

Hao Wang, Yixin Cao, Yangguang Li, Zhen Huang, Kun Wang, and Jing Shao. 2022. R2f: A general retrieval, reading and fusion framework for document-level natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3122–3134. The Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7208–7215. AAAI Press.

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.

Kilian Q Weinberger and Lawrence K Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. The Association for Computational Linguistics.

Linyi Yang, Shuibai Zhang, Zhuohao Yu, Guangsheng Bao, Yidong Wang, Jindong Wang, Ruochen Xu, Wei Ye, Xing Xie, Weizhu Chen, and 1 others. 2024. Supervised knowledge makes large language models better in-context learners. In *ICLR*.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024. Globesumm: A challenging benchmark towards unifying multi-lingual, cross-lingual and multi-document news summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10803–10821. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922. The Association for Computational Linguistics.

A Dataset Details

A.1 Premise Establishment Criteria

To ensure the quality and reliability of our CDCL-NLI dataset, we establish the following criteria for premise selection:

- **Content Parallelism**: The document pairs must discuss the same topic while being naturally written in their respective languages, rather than being translations of each other. This ensures authentic cross-lingual reasoning scenarios.
- **Information Complementarity**: While maintaining topic consistency, documents in different languages should present complementary perspectives or details, enabling meaningful cross-document inference tasks.
- **Language Distribution**: Premise document pairs are randomly sampled from different languages to reflect real-world cross-lingual scenarios. Each pair must consist of documents in two distinct languages, ensuring the dataset captures authentic cross-lingual reasoning challenges.

These criteria ensure that our dataset captures genuine cross-lingual reasoning challenges while maintaining natural language expression across different languages.

A.2 CDCL-NLI Label Definitions and Hypothesis Generation

Label Definitions. We define three inference labels for CDCL-NLI, considering various evidence distribution scenarios across documents:

- **Entailment**: The hypothesis is supported when either:
  - Evidence from both documents jointly supports the hypothesis through cross-document reasoning, or
  - One document provides sufficient supporting evidence while the other document contains no contradicting information

  In both cases, the conclusion must be logically derivable without requiring external knowledge.
- **Contradiction**: The hypothesis is contradicted when either:
  - Information from either document directly contradicts the hypothesis, or
  - The combined information from both documents leads to a logical conclusion that contradicts the hypothesis, or
  - The two documents present mutually contradictory evidence regarding the hypothesis
- **Neutral**: The relationship is neutral when:
  - Neither document alone nor their combination provides sufficient evidence to support or contradict the hypothesis, or
  - The documents contain only partially relevant information that doesn't allow for a definitive conclusion, or
  - The hypothesis introduces new information or claims that go beyond what can be verified from the documents

These definitions account for the complex nature of cross-document reasoning, where evidence may be distributed asymmetrically across documents and require different levels of information integration for reaching conclusions.

Hypothesis Creation. To generate high-quality hypotheses for our CDCL-NLI dataset, we designed a structured prompt for GPT-4o that specified detailed requirements for each label. The complete prompt template is reproduced in Figure 12.

This prompt design requires GPT-4o to generate evidence explaining the reasoning behind each hypothesis, which significantly reduces hallucination and improves alignment with the source documents. The structured output format facilitates automated processing while ensuring that each hypothesis is accompanied by clear justification of its entailment category. The generated hypotheses were subsequently reviewed by human annotators to ensure quality and adherence to the specified criteria.

A.3 Data Quality Assessment

Explanation Annotation Guidelines. We establish the following principles for EDU-based explanation annotation:

1. **Minimal Sufficiency**: Annotators should select the minimal set of EDUs that are necessary and sufficient to support the inference conclusion, avoiding redundant or irrelevant units.
2. **Cross-document Coverage**: Selected EDUs must include evidence from both premise documents when the inference requires cross-document reasoning, ensuring the explanation captures cross-lingual interactions.
3. **Logical Completeness**: The selected EDUs should form a complete logical chain that clearly demonstrates how the inference conclusion is reached.

Quality Metrics. We measured CDCL-NLI dataset using multiple metrics as shown in Table 6

The explanation component of our annotations was evaluated using three complementary metrics, all showing exceptional improvement after reconciliation:

- EDU Selection achieved 91% Jaccard similarity, indicating strong consensus on evidence selection
- Span Coverage reached 94% overlap ratio, demonstrating precise identification of relevant text spans
- Explanation Consistency achieved perfect alignment (1.00), ensuring logical coherence in reasoning

Our annotation quality assessment demonstrated strong reliability across all NLI categories. The inter-annotator agreement measured by Cohen's κ showed substantial initial agreement (0.71-0.75) and improved significantly after reconciliation (0.79-0.82). Specifically:

- Entailment labels achieved the highest final agreement ($\kappa = 0.82$)
- Contradiction cases showed strong consensus ($\kappa = 0.81$)
- Neutral instances, while slightly lower, maintained robust agreement ($\kappa = 0.79$)

Through our rigorous quality control and filtering process, we refined our dataset from an initial collection of 27,750 potential instances to 25,410 high-quality inference pairs. This 8.4% reduction reflects our commitment to maintaining high standards in both label accuracy and explanation quality, ensuring the dataset's reliability for both classification and interpretability research.

A.4 Data Information

Language Distribution. Figure 8 illustrates the language distribution of our dataset, where Spanish (15.3%), Russian (10.4%), and French (8.4%) represent the top three most frequent languages, while languages like Hebrew, Czech, and Hindi each accounts for approximately 1-2% of the data. This distribution not only reflects the imbalanced nature of multilingual usage in real-world scenarios but also ensures broad coverage of linguistic phenomena, enabling the study of diverse cross-lingual inference patterns.

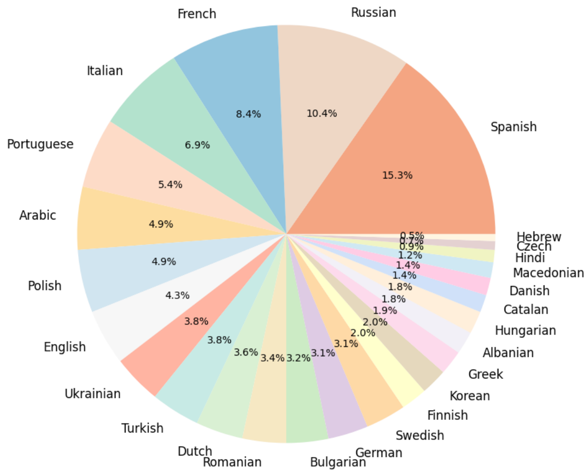


Figure 8: Language distribution of CDCL-NLI dataset.

Language Pair Distribution. As shown in Figure 9a, the dataset exhibits diverse language combinations across 24 languages. Spanish demonstrates the highest interaction frequency with other languages, particularly evident in Spanish-Russian (224 instances) and Spanish-Portuguese (178 instances) pairs. The heat map reveals several interesting patterns:

- Most language pairs maintain a balanced bidirectional relationship, with similar instance counts

Category	Description (Metric)	Init.	Final
NLI Label	Entailment (Cohen's κ)	0.75	0.82
	Neutral (Cohen's κ)	0.71	0.79
	Contradiction (Cohen's κ)	0.74	0.81
Explanation	EDU Selection (Jaccard Sim.)	0.76	0.91
	Span Coverage (Overlap Ratio)	0.81	0.94
	Explanation Consistency (Align.)	0.85	1.00

Table 6: Dataset quality assessment results.

in both directions
- Romance languages (Spanish, French, Portuguese, Italian) show stronger interconnections
- Less-resourced languages like Albanian and Macedonian have fewer cross-lingual pairs
- Russian and Spanish serve as central hub languages, connecting with most other languages in the dataset

EDU Count Distribution by Language Pair. The violin plot in Figure 9b illustrates the distribution of Elementary Discourse Units (EDUs) across the top language pairs. Several key observations emerge:
- Most language pairs show a median EDU count between 80 and 120 units
- The distributions are generally symmetric, indicating consistent EDU patterns regardless of the source language
- Romance language pairs (Romanian-Spanish, Portuguese-Spanish, Italian-Spanish) exhibit similar EDU distribution patterns
- Some pairs, particularly those involving Spanish as one of the languages, show wider distributions, suggesting more diverse discourse structures
- The violin shapes indicate that extreme EDU counts (very low or very high) are relatively rare across all language pairs

This analysis suggests that while the dataset maintains diverse language coverage, it also preserves consistent discourse complexity across different language combinations.

B Graph Construction Details

B.1 Relation Types

RST Graph Construction with Selected Relation Types. In constructing individual RST graphs for each document, we select a subset of relation types to focus on the most salient discourse and semantic connections. Specifically, we use the following relation types: Temporal, Summary, Condition, Contrast, Cause, Background, Elaboration, Explanation, and lexical chains. This selection balances coverage and complexity, ensuring that the resulting graph captures essential discourse relations and key semantic links without introducing excessive sparsity or noise. The inclusion of lexical chains further strengthens semantic cohesion by linking related words and expressions across different segments.

Graph Fusion with Extended Relation Types. During the fusion of RST graphs from multiple documents, we expand the set of relation types to include a broader range of discourse and organizational structures. The extended set comprises: Temporal, TextualOrganization, Joint, Topic-Comment, Comparison, Condition, Contrast, Evaluation, Topic-Change, Summary, Manner-Means, Attribution, Cause, Background, Enablement, Explanation, Same-Unit, Elaboration, and Lexical chains. This comprehensive set allows for richer cross-document alignment by capturing diverse forms of rhetorical and semantic relationships. Both in single-document and fused graphs, these relations serve as edge types in the construction of the Relation-aware Graph Attention Network (RGAT), enabling the model to effectively encode complex discourse and semantic structures.

B.2 Node Feature Definition

Specifically, for leaf nodes, we define:

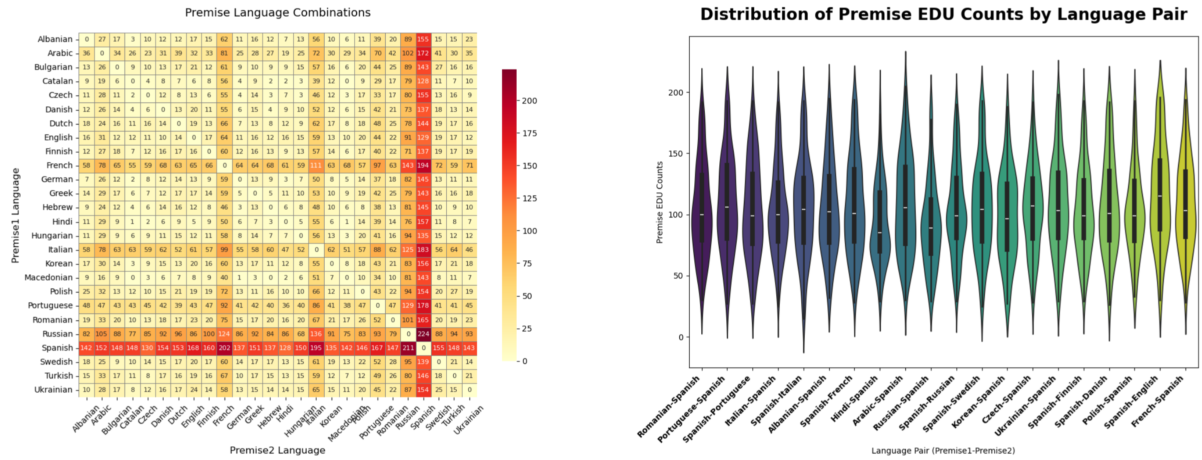
$$\phi(v_i) = \phi(\text{EDU}_s), \text{Text}_{v_i} = \text{EDU}_s, \text{Type}_{v_i} = 1.$$

For branch nodes, we define:

$$\phi(v_i) = \frac{1}{2}(\phi(v_j) + \phi(v_k)),$$

$$\text{Text}_{v_i} = \text{Text}_{v_j} \oplus \text{Text}_{v_k}, \text{Type}_{v_i} = 0,$$

where v_j, v_k are the children of v_i, and $\oplus$ denotes concatenation. For completeness, we provide the



(a) Heat map of premise language combinations across the dataset.

(b) Distributions of EDU counts across top-20 language pairs.

Figure 9: Statistic visualization of language pair distributions and their EDU characteristics.

detailed formulas for the relation-level and node-level attention mechanisms used in updating node embeddings.

B.3 Justification of the Cross-Document Edge Threshold δ

The threshold δ for adding cross-document lexical edges is set to 0.8 based on empirical analysis balancing sparsity and relevance of edges. We evaluated different threshold values on a validation set using the following metrics:

- **Edge Sparsity**: Higher thresholds reduce the number of edges, leading to sparser graphs that help avoid noise.
- **Semantic Relevance**: Lower thresholds introduce more edges but may include irrelevant or weakly related node pairs.
- **Downstream Task Performance**: We observed that $\delta = 0.8$ achieves the best trade-off, maximizing performance on the target task (e.g., accuracy or F1 score).

Figure 10 shows the impact of varying δ on edge count and task performance, confirming the choice of 0.8 as a reasonable and effective threshold.

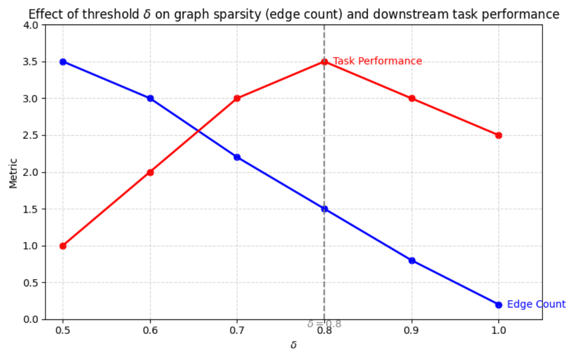


Figure 10: Effect of threshold δ on graph sparsity and task performance. Edge count (blue) decreases as δ increases, while task performance (red) peaks at $\delta = 0.8$ (dashed line), providing optimal balance between relevant connections and noise reduction.

Hyperparameters. For the model defined in Equation 1, the following settings are used: The first layer uses $K = 4$ attention heads. The second layer uses $K = 1$ attention head. Residual connections and dropout with rate 0.1 are applied after each layer.

B.4 Graph Attention Formulas

Relation Weight. The relation importance weights α_r are learnable parameters normalized by softmax:

$$\alpha_r = \frac{\exp(w_r)}{\sum_{r' \in R} \exp(w_{r'})},$$

where w_r is a trainable scalar parameter for rhetorical relation r.

Node-level Attention Coefficients. The attention coefficients $\beta_{ij,k}^{r,(l)}$ measure the importance of neighbor node v_j to node v_i under relation r, head k, and layer l. They are computed as:

$$\beta_{ij,k}^{r,(l)} = \frac{\exp\left(\psi\left(a_{r,k}^{(l)\top}\left[\mathbf{W}_{r,k}\mathbf{h}_{v_i}^{(l-1)} \| \mathbf{W}_{r,k}\mathbf{h}_{v_j}^{(l-1)}\right]\right)\right)}{\sum\limits_{v_m \in \mathcal{N}_r(v_i)} \exp\left(\psi\left(a_{r,k}^{(l)\top}\left[\mathbf{W}_{r,k}\mathbf{h}_{v_i}^{(l-1)} \| \mathbf{W}_{r,k}\mathbf{h}_{v_m}^{(l-1)}\right]\right)\right)}, \tag{8}$$

where $\mathbf{W}_{r,k}$ is the trainable linear transformation matrix for relation r and head k, $a_{r,k}^{(l)}$ is the learnable attention vector for relation r, head k, and layer l, $[\cdot\|\cdot]$ denotes vector concatenation, $\psi(\cdot)$ is the ELU activation function.

Additional Details. Each layer uses residual connections and dropout with a rate of 0.1 to improve training stability. The first layer uses $K = 4$ attention heads, while the second layer uses $K = 1$.

C Training Details

C.1 Model Training Hyperparameters

All the models are implemented in PyTorch and trained on an NVIDIA A100 GPU. To ensure fair comparison and reproducibility of results, all conventional baseline models and our model were fine-tuned under consistent experimental settings. As detailed in Table 7, each baseline utilizes the **XLM-RoBERTa-large** pretrained model as the base architecture and the **AdamW** optimizer for training. The learning rates are carefully selected for each model variant to optimize performance, while maintaining a uniform batch size of 16, a maximum input sequence length of 512 tokens, and training for 20 epochs. These standardized hyperparameters guarantee that performance differences stem from model design rather than training discrepancies, thereby supporting the validity and reproducibility of our comparative evaluation. Specially, for our model, as we split the documents into EDUs, so the maximux length is for one single EDU. By processing shorter EDUs instead of full documents, our model in long-text scenarios minimizes information loss, leading to improved performance.

C.2 LLM Fine-tuning Hyperparameters

For fine-tuning the Llama3-8B-instruct and Qwen3-8B model, we employed LoRA (Low-Rank Adaptation) to efficiently adapt the large-scale pretrained model with limited computational resources. The key hyperparameters for LoRA tuning included a rank of 16, which balances adaptation capacity and parameter efficiency, and a dropout rate of 0.1 to mitigate overfitting. The learning rate was set to 2×10^{-4} with a linear warmup over the first 500 steps, followed by a constant decay. We used a batch size of 64 sequences and capped the maximum input length at 1024 tokens to fully leverage the model's context window. Training was conducted for 10 epochs, which empirically provided a good trade-off between convergence and training cost. These hyperparameters were chosen based on prior LoRA tuning best practices and preliminary experiments to ensure stable and effective adaptation of the Llama3-8B-instruct and Qwen3-8B model. The prompt is shown in Figure 11.

D Additional Experiments

D.1 LLM Few-shot Prompt

As shown in Figure13, one example is provided to demonstrate how to determine the logical relationship between the premise and the hypothesis. The model is instructed to output exactly one of three labels: *entailment*, *contradiction*, or *neutral*. This prompt effectively guides the model to understand the task objective and output format, thereby enhancing its reasoning capability across multiple languages and documents during the few-shot validation stage.

D.2 LLM in Zero-shot Scenario

The zero-shot results reported in Table 8 are obtained using the same prompt design as the few-shot experiments, differing only in the absence of in-context examples. As expected, all models perform worse under the zero-shot setting compared to their few-shot counterparts, demonstrating the effectiveness and necessity of providing exemplars in the prompt for this task. Despite the overall performance drop, the relative ranking of the three models remains consistent with the few-shot scenario, with Qwen3-8B achieving the highest scores, followed by GPT-4o, and then Llama-3-8B. This consistency indicates that these models' capabilities in handling the CDCL-NLI task are stable across different prompting strategies. Moreover, the results highlight the challenge of zero-shot cross-document and cross-lingual natural language inference, emphasizing the importance of prompt engineering and in-context learning to boost model performance on complex multilingual and multi-document reasoning tasks.

D.3 Baseline Evaluation in Single Document Scenario

To further demonstrate the cross-document characteristic of our dataset, we add this extra experiment to evaluate the performance using either a single document ($Document_1$ or $Document_2$) as the premise compared to using the full combined premise, as summarized in Table 9. The noticeable improvement in F1 score when both documents are combined indicates that effective inference relies on integrating information from multi-

Baseline	Base Model	Optimizer	LR	Batch Size	Max Length	Epochs
Hypothesis-only	XLM-R Large	AdamW	3×10^{-6}	16	512	20
DocNLI	XLM-R Large	AdamW	3×10^{-6}	16	512	20
R2F	XLM-R Large	AdamW	1×10^{-6}	16	512	20
Ours	XLM-R Large	AdamW	1×10^{-5}	16	512(per EDU)	20

Table 7: Training hyperparameters for conventional baseline models and our model. These configurations, including the consistent use of the XLM-RoBERTa-Large base model and AdamW optimizer, were utilized to ensure reproducibility and fair comparison.

Fine-tunning Prompt

You are skilled in the NLI task. Given a premise consisting of two documents and a hypothesis, each with its specified language, your task is to determine the natural language inference (NLI) relationship between the hypothesis and the premise. Note that the premise and hypothesis may be in different languages. The output should be one of three labels: `Entailment`, `Contradiction`, or `Neutral`.

Input format:
Premise 1 (Language: <Lang1>): <Premise1 text>
Premise 2 (Language: <Lang2>): <Premise2 text>
Hypothesis: <Hypothesis text>
Output format:
One of the labels: `Entailment`, `Contradiction`, or `Neutral`
—
Example:
Premise 1 (Language: English): The cat is sitting on the mat.
Premise 2 (Language: French): Le chat est assis sur le tapis.
Hypothesis: The animal is resting on a rug.
Output: Entailment
—
Now, given the input premises and hypothesis, provide the NLI label.

Figure 11: Llama3-8B-Instruct and Qwen3-8B Finetuning Prompt.

ple sources. Additionally, the similar results observed between **Single Document 1** (54.22% and 57.09% F1) and **Single Document 2** (54.95% and 57.12% F1) imply that each document provides valuable and roughly equal contributions. This further supports the notion that reasoning in this task benefits from synthesizing evidence across documents rather than focusing on a single source.

E Case Study

E.1 Our Method Case

Our approach employs a multi-stage framework for analyzing complex multi-document multi-lingual NLI scenarios. Take the given example in Figure 13, the Yanaquihua gold mine incident in Condesuyos, Peru, where a short circuit-induced fire resulted in 27 fatalities among workers trapped within a tunnel, prompting mobilization of local authorities and rescue teams. We begin by parsing the premise documents using Rhetorical Structure Theory (RST), which generates hierarchical discourse trees wherein each node represents an Elementary Discourse Unit (EDU). These nodes are assigned unique indices, with their textual content comprehensively documented in Tables 10 and 11.

Following RST parsing, we construct individual discourse graphs for each premise document. These discrete graphs are subsequently integrated into a unified premise graph through the establishment of "Lexical" chains that leverage semantic information and discourse relations to facilitate enhanced inference. As illustrated in Tables 10 and 11, EDU nodes sharing identical uppercase character designations indicate the presence of cross-document “Lexical” chains. This consolidated graph representation effectively captures

Model	TestSet1: Cross-Lingual			TestSet2: English		
	Precision	Recall	F1 Macro	Precision	Recall	F1 Macro
Llama-3-8B	44.00	50.00	46.00	49.00	55.00	50.00
GPT-4o	50.00	54.00	52.00	59.00	62.00	61.00
Qwen3-8B	58.00	54.00	57.00	68.00	64.00	63.00

Table 8: Zero-shot performance of large language models on the CDCL-NLI dataset.

Model	Single Document1	Single Document2	Combined Documents
DocNLI	54.22	54.95	64.46
R2F	57.09	57.12	65.42

Table 9: F1 Macro scores for different methods across premises with varying numbers of documents.

EDU	Text	EDU	Text
1	7. května	22	řekl prokurátor Giovanni Matos místní televizní stanici Canal N.
4	Společnost okamžitě nereagovala na žádost o komentář.	24	jsou 27 obětí,"
7	(Reuters) -	25	„Informace jsou správné,
11①	Úřadníci uvedli v neděli, že nehoda v malé zlaté dolině na jihu Peru odnesla život 27 pracovníků.	26	potvrdila je policie v Yanaquihuě,
12	Jedná se o jeden z nejúmrtnějších důležitých událostí v těžebním průmyslu v tomto jihoamerickém státě.	27	„Jedná se o formální dolinu (...),
15②	Nehoda se stala v sobotu ráno v těžební společnosti Yanaquihua, která se nachází v provincii Condesuyos v departementu Arequipa.	30	dodal.
17	Zdá se, že došlo ke zkratu, která způsobila požár uvnitř tunelu,	33	musíme jít
18	uvedla regionální vláda.	34	a zjistit, kde jsou mrtví, jestli je tam bezpečné,
37③	Regionální vláda Arequipy a ministerstvo vnitra mobilizovaly policie, zdravotníky a sanitky, aby pomohly při péči o oběti a jejich záchraně.	35	aby se tam mohli dostat policisté a soudní pracovníci
39	Podle statistik peruánského ministerstva těžeb a energie je toto nejvyšší počet obětí v jediném těžebním nehodě	36	a provést procedury,"
40	nejméně od roku 2000.		

Table 10: Elementary Discourse Units (EDUs) from $Document_1$ with their corresponding Spanish text. Segments highlighted in green represent evidence supporting the Entailment classification. EDU indexes with circled numbers ① indicate cross-document "Lexical" chains linking to corresponding EDUs in $Document_2$.

the comprehensive discourse context across the premises, enabling more robust and coherent semantic modeling.

The classification module processes this unified graph in conjunction with the hypothesis to predict the appropriate NLI label. Concurrently, the explanation extraction module identifies a salient subset of nodes within the premise graph that substantiate the classification decision. These explanation nodes are visually distinguished through green font highlighting in Tables 10 and 11, explicitly denoting their explanatory significance.

Our integrated methodology capitalizes on the hierarchical discourse structure inherent in RST parsing and the semantic connectivity across documents, ensuring that the model's inference is both accurate and interpretable. The explicit identification of explanation nodes within the discourse structure facilitates transparent, human-comprehensible rationales grounded in the premise texts, thereby advancing the explainability of NLI systems in complex multi-document, multi-lingual scenarios. This approach proves particularly valuable when analyzing intricate real-world situations such as the Yanaquihua mine disaster, where understanding the causal relationships and contextual factors is crucial for proper inference.

EDU	Text	EDU	Text
14	informó el Ministerio Público de ese país.	53	[Al menos siete muertos en Texas
15①	Al menos 27 personas murieron en Perú	54	tras atropellamiento en una parada de autobús cerca de un refugio para inmigrantes]
17	y otras dos fueron rescatadas	56	lo que impidió que los mineros pudieran escapar.
18	luego de un incendio el sábado en una mina de oro en la sureña provincia de Condesuyos,	57	Se informó que
21	Según las primeras investigaciones, la tragedia tuvo lugar	59	el fuego se propagó de manera muy rápida por las estructuras de madera que sostienen el yacimiento,
23②	tras producirse un cortocircuito a 100 metros de la entrada de la mina Yanaquihua,	60	dedicado a la extracción de oro,
24	conocida como Esperanza I.	61	Medios locales peruanos indicaron que
28	informó el Gobierno regional de Arequipa.	63	27 trabajadores quedaron atrapados en la mina
29	“Se habría producido un cortocircuito	64	tras un incendio.
31	que provocó un incendio en el interior del socavón,	65	Getty Images
32	que habría puesto en riesgo la vida de los trabajadores”,	71	James Casquino, alcalde de Yanaquihua, dijo que
33	Medios locales indicaron que	73	el dueño de la mina fue a la comisaría de ese distrito
34	27 trabajadores atrapados habían fallecido por asfixia.	75	para pedir ayuda en el rescate de las personas
35	La noche del sábado, el Ministerio del Interior confirmó en su cuenta de Twitter el accidente.	76	que se encontraban atrapadas.
38	indicó el tuit.	78	[Mueren varios migrantes en un accidente de auto en Nuevo México cerca de la frontera]
39	“Personal policial se encuentra en el distrito de Yanaquihua	79	Las autoridades indicaron que
41	para apoyar en las labores de rescate de los cuerpos de mineros	80③	hacia la zona se habían movilizado rescatistas.
42	que fallecieron dentro de un socavón en la provincia de Condesuyos”,	81	Familiares de las víctimas se reunieron frente a la comisaría de Yanaquihua
49	Imágenes difundidas en redes sociales mostraban una gran columna de humo negro proveniente de la mina,	83	para recabar información sobre la suerte de sus seres queridos
51	y medios locales indicaron que	84	y exigir a las autoridades que agilizaran las labores de rescate de los cuerpos.
52	en el momento del cortocircuito había personal trabajando a unos 80 metros de profundidad.	85	El fiscal Giovanni Matos indicó a un medio local que
87	las tareas en la mina podían demorar	89	porque no se sabía si los equipos de rescatistas podían ingresar a la mina
23	para retirar los cadáveres.	90	para retirar los cadáveres.
91	[Una tormenta de polvo en Illinois causa múltiples muertes y decenas de hospitalizados tras choque masivo]	94	indica la compañía en su página web.
95	La mina pertenece a Yanaquihua S. A. C., una empresa	96	que reúne a pequeños productores mineros dedicados a la explotación del oro y otros metales,

Table 11: Elementary Discourse Units (EDUs) from $Document_2$ with their corresponding Spanish text. Segments highlighted in green represent evidence supporting the Entailment classification. EDU indexes with circled numbers ① indicate cross-document "Lexical" chains linking to corresponding EDUs in $Document_1$.

E.2 LLM Answer Case

As shown in Table 3, Qwen3-8B achieves higher scores compared to Llama3-8B-instruct and the closed-source GPT-4o. One key reason is that we evaluate Qwen3-8B using its `thinking` (chain-of-thought) mode, as illustrated in Figure 14. We still take the case in validation prompt(Tabel 13) as an example, the model systematically parses each premise, accurately extracts key facts, and performs detailed cross-checking between the articles and the hypothesis. It also demonstrates the ability to handle subtle differences in wording (such as distinguishing between deaths and rescues) and to resolve potential ambiguities in translation (e.g., the meaning of "obětí" in Czech).

Nevertheless, our proposed approach still outperforms Qwen3-8B, primarily due to its ability to explicitly capture document structure through RST parsing and cross-document, cross-lingual semantic integration via "Lexical" chains. Moreover, our method demonstrates superior efficiency with significantly lower computational requirements and faster inference time, making it more practical for real-world applications while maintaining state-of-the-art performance.

Hypothesis Generation Prompt

[Hypothesis Generation Prompt] We are creating a cross-document cross-lingual NLI dataset. Below are two documents under the event topic: [CATEGORY], treated as one premise in this NLI task. Based on them, generate hypotheses in three labels. You must **strictly follow** the instructions:
1. Hypothesis: The hypothesis should be a factual statement based on the content of the articles. It must be a simple statement and **should not contain any explanation or analysis like "this contradicts" or "this agrees with" or "this is inconsistent with."**
2. Evidence: The evidence section should explain how the hypothesis relates to the articles, including any contradictions or confirmations, using specific quotes from the articles.
Document Details:
- **Document 1**: Date: [DATE_1]; Article: [ARTICLE_1]
- **Document 2**: Date: [DATE_2]; Article: [ARTICLE_2]

[Task 1: Entailment Generation] Generate an Entailment Hypothesis and evidence.
The hypothesis is supported if evidence from both documents together or from one document alone (without contradiction in the other) logically supports it.
Guidelines:
- Ensure each detail is verifiable by premise
- Include specific facts (dates, names, etc.)
- No speculation—strictly based on facts

Evidence:
- Quote relevant parts from both articles and explain how they jointly support the hypothesis

[Task 2: Neutral Generation] Generate a Neutral Hypothesis and evidence.
One hypothesis is neutral if there is insufficient or only partial evidence in the premise to confirm or deny it, or if it contains information beyond what the premise verify.
Guidelines:
- Reasonable speculation or expanded related aspects in a reasonable way
- Propose middle ground if there's conflicting information

Evidence:
- Show partial support from one or both articles without full confirmation
- Explain how the hypothesis goes beyond but stays consistent with the Document content

Remember, A neutral hypothesis should not be directly confirmed by the premise (which would make it entailed), nor should it contradict the articles (which would make it conflicting).
[Task 3: Conflicting Generation] Generate a Conflicting Hypothesis and evidence.
One hypothesis is contradicted if either document or their combined information directly opposes it, or if the documents conflict with each other regarding the hypothesis.
Guidelines:
- Negate or reverse key information in premise
- Complex and multi-faceted hypothesis with multiple contradictions
- Try to combine multiple points of contradiction
- Ensure the hypothesis appears reasonable but actually conflicts clearly

Evidence:
- Show which document(s) the hypothesis contradicts and explain specific points
- If applicable, explain why this hypothesis cannot coexist with the premise content

Output in JSON format:

```
{ "entail_evidence": "...",
  "entail_hypothesis": "...",
  "neutral_evidence": "...",
  "neutral_hypothesis": "...",
  "conflict_evidence": "...",
  "conflict_hypothesis": "..."}
```

Figure 12: Hypothesises Generation Prompt.

Validation Prompt

You are tasked with a cross-document and cross-language Natural Language Inference (NLI) task. Your goal is to determine the relationship between the "premise" and the "hypothesis". The premise consists of two documents presented in different languages. Here is one example:

Premise(Document1 in Spanish)

Al menos 27 personas murieron en Perú y otras dos fueron rescatadas luego de un incendio el sábado en una mina de oro en la sureña provincia de Condesuyos, informó el Ministerio Público de ese país. Según las primeras investigaciones, la tragedia tuvo lugar tras producirse un cortocircuito a 100 metros de la entrada de la mina Yanaquihua, conocida como Esperanza I. “Se habría producido un cortocircuito que provocó un incendio en el interior del socavón, que habría puesto en riesgo la vida de los trabajadores”, informó el Gobierno regional de Arequipa. Medios locales indicaron que 27 trabajadores atrapados habían fallecido por asfixia. La noche del sábado, el Ministerio del Interior confirmó en su cuenta de Twitter el accidente. Personal policial se encuentra en el distrito de Yanaquihua para apoyar en las labores de rescate de los cuerpos de mineros que fallecieron dentro de un socavón en la provincia de Condesuyos. Imágenes difundidas en redes sociales mostraban una gran columna de humo negro proveniente de la mina, y medios locales indicaron que en el momento del cortocircuito había personal trabajando a unos 80 metros de profundidad. Se informó que el fuego se propagó de manera muy rápida por las estructuras de madera que sostienen el yacimiento, dedicado a la extracción de oro, lo que impidió que los mineros pudieran escapar.James Casquino, alcalde de Yanaquihua, dijo que el dueño de la mina fue a la comisaría de ese distrito para pedir ayuda en el rescate de las personas que se encontraban atrapadas.

Premise(Document2 in Czech)

7. května (Reuters) - Úřadníci uvedli v neděli, že nehoda v malé zlaté dolině na jihu Peru odnesla život 27 pracovníků. Jedná se o jeden z nejúmrtnější důležitých událostí v těžebním průmyslu v tomto jihoamerickém státě. Nehoda se stala v sobotu ráno v těžební společnosti Yanaquihua, která se nachází v provincii Condesuyos v departementu Arequipa. Zdá se, že došlo ke zkratu, která způsobila požár uvnitř tunelu, uvedla regionální vláda. „Informace jsou správné, potvrdila je policie v Yanaquihuě, jsou 27 obětí,“ řekl prokurátor Giovanni Matos místní televizní stanici Canal N. Regionální vláda Arequipy a ministerstvo vnitra mobilizovaly policii, zdravotníky a sanitky, aby pomohly při péči o oběti a jejich záchraně. Podle statistik peruánského ministerstva těžeb a energie je toto nejvyšší počet obětí v jediném těžebním nehodě nejméně od roku 2000. Společnost okamžitě nereagovala na žádost o komentář.

Hypothesis and Label

Hypothesis: "The Yanaquihua gold mine in the province of Condesuyos, Peru, experienced a fatal incident on a Saturday morning due to a short circuit that caused a fire, leading to the death of 27 workers trapped inside the tunnel, with local authorities and rescue teams being mobilized to the site."
Output: entailment

Now, given the following input, output exactly one word indicating the NLI label: `neutral`, `entailment`, or `contradiction`.
Premise:[Premsise]
Hypothesis: [Hypothesis]
Output the NLI label (neutral/entailment/contradiction):

Figure 13: LLM Validation Prompt.

Qwen3-8B Reasoning Output

Okay, let's tackle this NLI task. So, the premise has two articles in Spanish and Czech, and the hypothesis is in English. I need to determine if the hypothesis is entailed, contradicted, or neutral based on the premises.

First, I'll read through the Spanish article. It mentions that on Saturday, there was a fire in a gold mine in Condesuyos province, caused by a short circuit 100 meters from the entrance. The mine is called Yanaquihua, known as Esperanza I. They say 27 people died, two were rescued. The regional government of Arequipa mentioned the short circuit caused the fire, risking workers' lives. Local media said 27 trapped workers died of asphyxia. The Ministry of Interior confirmed the accident on Saturday, and police were there for rescue. The fire spread quickly through wooden structures, preventing escape. The mine belongs to Yanaquihua S.A.C., a company with small miners.

Now the Czech article: Officials said the accident in the small gold mine in southern Peru resulted in 27 deaths. It happened on Saturday morning at the Yanaquihua mine in Condesuyos province, Arequipa department. A short circuit caused a fire inside the tunnel. The regional government confirmed 27 victims. The prosecutor Giovanni Matos said the information is correct, and police are there to find the dead, check safety for rescue teams. Regional government and interior ministry mobilized police, medics, and ambulances. It's the highest number of deaths in a single mining accident since 2000. The company didn't respond to a comment request.

Hypothesis: The Yanaquihua gold mine in Condesuyos, Peru, had a fatal incident on Saturday morning due to a short circuit causing a fire, leading to 27 workers' deaths trapped inside the tunnel, with local authorities and rescue teams mobilized.

Check each part. The Spanish article says the fire was caused by a short circuit, leading to 27 deaths. The Czech article confirms the same: short circuit causing fire, 27 deaths. Both mention Saturday morning (the Spanish says "sábado" which is Saturday, and the Czech says "v sobotu ráno" which is Saturday morning). The mine is Yanaquihua in Condesuyos, Peru. The death count is 27, and both mention authorities and rescue teams being mobilized. The hypothesis states all these points. The Spanish article also mentions the mine's name and the company, which the Czech article doesn't, but that's not conflicting. The hypothesis doesn't mention the company, so that's okay.

Are there any contradictions? The Spanish article mentions two people rescued, but the hypothesis says 27 died. Wait, the hypothesis says "leading to the death of 27 workers trapped inside the tunnel." The Spanish article says 27 died and two were rescued. The Czech article says 27 victims, which could include the two rescued? Wait, maybe not. If two were rescued, then the deaths would be 27, and the two rescued are alive. So the hypothesis says 27 workers died, which matches the Spanish article's 27 deaths. The Czech article says "27 obětí" which is victims, which could be the 27 who died, but maybe the two rescued are not counted as victims? Or maybe the Czech article is using "obětí" as victims, which could include those who died and those rescued? Wait, in Czech, "obětí" is victims, usually meaning those who died. But the Spanish article says 27 died and two rescued. The Czech article says 27 victims, which might be the same 27, but the two rescued might not be considered victims. So the hypothesis says 27 workers died, which aligns with both articles. The Czech article says "27 obětí" which is victims, so that's consistent.

Also, the hypothesis mentions the fire caused by a short circuit, which both articles confirm. The mobilization of authorities and rescue teams is mentioned in both. The time is Saturday morning in both. The mine's name and location are correct.

Therefore, the hypothesis is fully supported by both articles. So the label should be entailed.

Figure 14: Illustration of Qwen3-8B's chain-of-thought reasoning process. The model systematically analyzes premise documents by extracting key facts, performing cross-document verification, and resolving translation ambiguities (e.g., Czech term "obětí"). While this reasoning approach enables Qwen3-8B to outperform other LLMs like Llama3-8B-instruct and GPT-4o, our RST-based method achieves superior performance through more intuitive explanations, explicit discourse structure modeling, and cross-lingual semantic integration, all while requiring significantly lower computational resources and faster processing times.