PATeam at SemEval-2025 Task 10: Two-stage News Analytical Framework: Target-oriented Semantic Segmentation and Sequence Generation LLMs for Cross-Lingual Entity and Narrative Analysis

Ling Sun **Xue Wan** **Yuyang Lin** **Fengping Su** **Pengfei Chen**
Ping An Life Insurance Company of China, Ltd.
sunling583@163.com wx18707735705@163.com lyy476629663@gmail.com
fengpings@outlook.com 1012673739@qq.com

Abstract

This paper presents our approaches for three subtasks in SemEval-2025 Task 10, which focus on entity framing, narrative classification, and narrative extraction in news analysis, respectively. We propose a two-stage news analytical framework for both Subtask 1 and 2. In Subtask 1 (Entity Framing), we design an entity-oriented data processing pipeline to address the issue of redundant information in a news article, and explore ways to use multilingual datasets effectively through sufficient experiments. The system achieves the first place in Bulgarian and the second place in English and Portuguese. In Subtask 2 (Narrative Classification), a similar narrative-oriented data processing pipeline is adopted to obtain condensed news chunks for each narrative. We conduct in-depth discussion regarding approaches to enhancing both data quality and volume, and explore one-vs-rest classification models and sequence prediction models for multi-label classification tasks. The system ranks first in Bulgarian and second in Russian and Portuguese. In Subtask 3 (Narrative Extraction), we build our system with data augmentation, supervised fine-tuning, and preference-based reinforcement learning. This system achieves the first place in Bulgarian, Russian and Hindi and the second place in Portuguese.

1 Introduction

The rise of the Internet and artificial intelligence has revolutionized how we access information, but also leaves us vulnerable to manipulative content and disinformation. Various tasks on media analysis have studied entity roles in memes (Barrón-Cedeño et al., 2024) and persuasion techniques in textual and multimodal datasets (Piskorski et al., 2023; Dimitrov et al., 2021). SemEval-2025 Task 10 (Piskorski et al., 2025; Stefanovitch et al., 2025), building on top of these tasks, proposes the challenge of developing cutting-edge NLP systems for multilingual characterization and extraction of narratives from online news. The task aims to automatically identify narratives and the roles of the relevant entities involved. These news analytics capabilities are essential for studying disinformation phenomena on specific targets.

The task focuses on three core news analytics challenges in five languages (Bulgarian, English, Hindi, Portuguese, and Russian).

- Entity Framing: Determine how entities are portrayed.
- Narrative Classification: Identify all the storylines of a news article.
- Narrative Extraction: Generate short explanations for dominant narratives.

The original dataset poses two main challenges. Firstly, Training classification models with entire articles leads to poor performance. We observe that new articles usually contain multiple entities and narratives and, therefore, redundant or interfering information for identifying a specific target.

The small number of articles compared to the number of different labels and label power sets is another challenge. In cases where there is not enough supervised data to train a BERT-like model, the advantages of leveraging the pre-trained knowledge of LLMs to solve downstream tasks with a small amount of data become apparent.

In summary, this paper presents the following contributions:

- Target-oriented semantic segmentation with multiple prompts is adopted to alleviate redundant information within news articles and leads to significant performance gain.
- Sufficient experiments with multilingual and monolingual approaches are conducted to identify effective ways to utilize cross-lingual datasets.
- The correlation between coarse-grained and fine-grained roles or narratives is captured by multi-turn dialogues training set for LLMs, which further improves model performance.
- Finally, our systems achieve five first places, five second places, and several top-10 rankings on SemEval 2025 task 10 Test Leaderboards.

We present specific approaches and observations for each subtask separately in Sections 4 to 6.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2121–2132

2 Related Work

(Zhou et al., 2024) addresses the problem of noisy information by employing language models to automatically generate data cleaning programs and developing manual rules to enhance the quality of pre-training data. This method improves corpus usability with semantic segmentation techniques, such as filtering redundant paragraphs and removing low-quality texts. Similarly, (Gehman et al., 2020) utilizes semantic segmentation to identify toxic content in generated texts, combining rule-based and model-based predictions to filter harmful segments. Building upon these approaches, we leverage the ability of open-source LLMs and propose a target-oriented semantic segmentation technique during the data preparation phase and achieve significant improvements.

To address the challenge of low-resource scenarios in NLP tasks, Wei and Zou propose data augmentation methods including Synonym Replacement, Random Insertion, Random Swap, and Random Deletion (Wei and Zou, 2019). Advanced strategies namely metadata-aware data augmentation, which exchange similar products within a food category (Zhang et al., 2021), and prototypical networks, (Snell et al., 2017) show promise for data augmentation in professional domains. To take full advantage of the world knowledge within open-source LLMs, including LLaMa3.3-70B (AI@Meta, 2024) and Qwen2.5-72B(Yang et al., 2024), we design various prompt to generate multiple sets of training data to mitigate the problem of unbalanced data and overfitting in low-resource scenarios. (Wang et al., 2023) discusses multilingual training as another solution to low-resource scenarios of African languages. We perform sufficient experiments on combinations of monolingual datasets in the context this task and identify effective ways to conduct multilingual training.

Multi-dimensional Type-slot Label Interaction Network (MTLN) proposed by (Wan et al., 2023) is a neural network designed to address multiple NLP tasks using a unified architecture. Compared with single-task learning, multi-task learning (MTL) shows enhanced generalization abilities by learning task correlations and complementary features, and mitigates the issue of underfitting (Guo et al., 2018). Built upon this idea, we transform the correlation between coarse-grained and fine-grained roles or narratives into a multi-turn dialogue to build multi-task training sets for LLMs and observe a steady increase.

3 System Overview

3.1 Model Training Approaches

3.1.1 Supervised Fine-Tuning (SFT)

Supervised Fine-Tuning (SFT) is the primary method used in three subtasks. During the SFT stage, model weights are updated by minimizing a supervised loss function to encourage better predictions of labeled data. The loss function during SFT is generally defined as:

$$\mathcal{L}_{SFT} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{cross-entropy}}(y_i, \hat{y}_i) \tag{1}$$

where $\mathcal{L}_{\text{cross-entropy}}$ is the cross-entropy loss of the predicted label sequence $\hat{y}_i$ given the true label sequence y_i, and N is the number of training samples.

3.1.2 Reinforcement Learning (RL)

Reinforcement Learning (RL) is a common approach to improve model performance in hard cases where supervised fine-tuning fails. While the PPO (Proximal Policy Optimization) algorithm (Schulman et al., 2017) is widely used to align the behavior of pre-trained LLMs with complex human preferences, DPO (Direct Preference Optimization) (Rafailov et al., 2024) shows decent performance in downstream tasks with simple goals or evaluation metrics such as classification and summarization. The DPO algorithm also requires much less computational resources by transforming an RL problem into an SFT problem. The objective of DPO is to maximize the probability ratio of selecting a referred output over a rejected output. The objective function is written as:

$$\begin{aligned}\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = \\ - E_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \\ - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big]\end{aligned} \tag{2}$$

where π_θ is the policy model to be trained, π_{ref} the reference model that is kept frozen, y_w the chosen sequence, and y_l the rejected sequence.

3.1.3 LoRA (Low-Rank Adaptation)

To implement SFT and RL in low-resource scenarios, LoRA (Low-Rank Adaptation)(Hu et al., 2022) is adopted for parameter-efficient tuning. The idea of LoRA is to approximate model weight updates ΔW by the product of two low-rank matrices A and B, reducing the number of parameters to be trained while maintaining model performance. The process can be written as:

$$W_{\text{new}} = W + \Delta W = W + AB^T \tag{3}$$

where A and B are two trainable matrices and ΔW is kept frozen to improve computational efficiency. The loss function for LoRA is typically the standard SFT loss.

3.2 Data Processing Techniques

3.2.1 Data Preprocessing

Both the experimental results and the text analysis show that it is not an effective approach to use entire news articles as input to train classification models. An excessively long input contains redundant or noise information. In Subtask 1, a news article may contain multiple

entities and one entity (such as Russia) may also appear multiple times, each with a different fine-grained role. The mixture of different targets in an article makes it difficult for the model to accurately identify the characteristics of each one. The case is similar in Subtask 2. The latest advances in generative LLMs provide a convenient solution. We build prompts to describe each entity and narrative and use LLMs for semantic segmentation to obtain the most relevant contextual information for each target.

3.2.2 Data Augmentation

(Mahmoud et al., 2025) presents a corpus of 1,378 recent news articles in five languages (Bulgarian, English, Hindi, Portuguese and Russian). However, it is difficult to train a monolingual model with robust representations with only hundreds of examples in each language. To address data sparsity, we perform data augmentation by rewriting, summarizing, and semantic segmentation with multiple prompts, and resampling to improve data quality and alleviate unbalanced distribution across languages. We also investigate using multilingual training data as a technique for data augmentation as well.

3.3 Multi-label Classification

Multi-label classification is a machine learning task that addresses problems where instances are associated with multiple interdependent labels. Traditional approaches often transform the problem into binary subtasks (e.g., one-vs-rest, one-vs-one). These methods may struggle with label correlations and scalability when the label space is large. We adopt the method of label sequence generation to model interdependent relationships among labels and expand this idea by leveraging the ability of multiturn conversations of LLMs to model hierarchical relationships. The performance of these systems are further improved by constructing high-quality training data.

3.4 Ensemble Learning for Performance Boosting

Voting with multiple models is an effective approach to correct obvious errors that occur with a single model, and therefore, reduces model variance and improves generalization of a system (Ruta and Gabrys, 2005; Zhang et al., 2014). We sample multiple subsets of training data and fine-tune LLMs with LoRA on each subset. The checkpoint with the best validation result is chosen for each fine-tuning process. Finally, we run inference with each chosen checkpoint and vote for a final output.

4 System Description for Subtask 1

4.1 Model Design

Subtask 1 is a hierarchical multi-label classification task of entity roles in news articles. As discussed in the previous sections, the challenge is redundant information in long articles and limited data for each language. To address the first challenge, we propose a two-stage news analytical framework to enhance the accuracy of entity role classification. The first stage is semantic segmentation. Only paragraphs or sentences that are closely related to the target entity are extracted to form a coherent context for classification. This step facilitates data cleaning and dynamically restructure the context according to the target entity. The second stage is hierarchical multi-label classification. Phi-4(Abdin et al., 2024) and Qwen2.5-32B(Yang et al., 2024) models are fine-tuned using LoRA to specialize in multi-turn conversational reasoning, enabling nuanced understanding of role-specific patterns at both coarse and fine through sequential interaction analysis.

For the second challenge, we build multiple prompts to rewrite, summarize, and extract context from the news text for data augmentation. We also conduct sufficient experiments to explore the use of multilingual training set in depth.

As we design four prompt-based data augmentation methods, five sets of training data are obtained in total. Consequently, we create a five-fold validation of LoRA fine-tuning of LLMs. The best checkpoint in each validation is selected, and the final best results are given by a majority voting mechanism. This ensemble method leverages the collective wisdom of multiple models, reducing the potential biases and errors of a single model, and enhancing the overall model performance and generalization ability.

4.2 Experimental Setup

We utilize LLama-Factory (Zheng et al., 2024) to implement training setups. LoRA is adopted to fine-tune Phi-4 and Qwen2.5-32B, with rank=4, alpha=8, and low-rank adapters applied to all layers. Adam optimization (Kingma and Ba, 2014) with a warm-up step of 10% and a learning rate of 1e-4. Distributed training is implemented with Deepspeed Zero-3 on two NVIDIA A100 GPUs (80GB) , with a batch size of 2 per device and a gradient accumulation of 32, training for a total of 3 epochs.

4.3 Best Results

The best results on the validation datasets are obtained by fine-tuned Qwen2.5-32B-Instruct for English and Portuguese and by Phi-4 for Russian, Bulgarian and Hindi. The target-oriented semantic segmentation technique leads to significant boost in model performance. The use of multilingual datasets and a voting mechanism bring additional benefits. The results are shown in Table 1.

Dataset	Language	Phi-4 EMR%	Qwen2.5 EMR%
EN+PT	English	53.65	**54.95**
EN+PT	Portuguese	76.99	**77.59**
EN+RU+BG	Russian	**65.12**	61.63
RU+BG	Bulgarian	**61.29**	54.48
HI	Hindi	**48.93**	44.64

Table 1: Best results on the validation set for Subtask 1

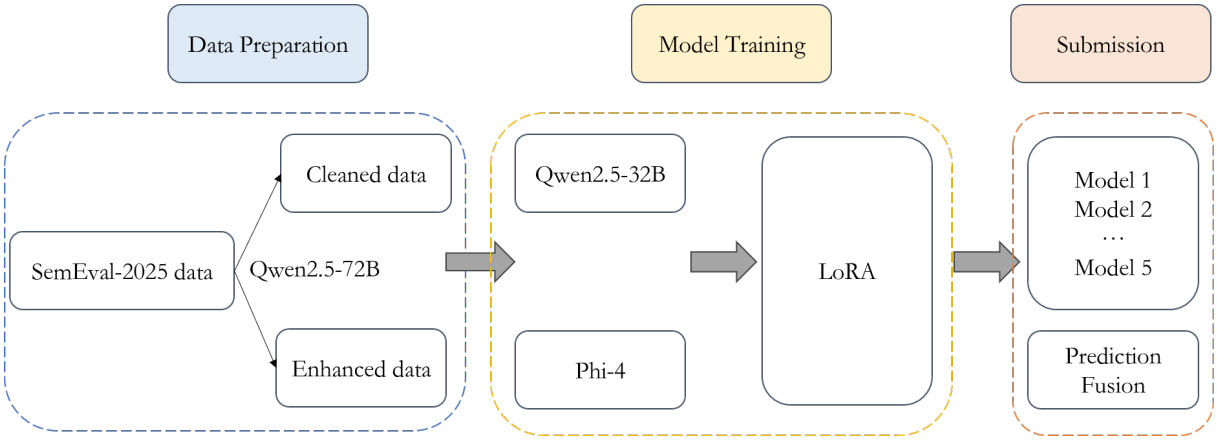

Figure 1: Task experimental progress

we select the best-performing model settings for each language and conduct inference on the test set. As presented in Table 2, the system achieves the first place in Bulgarian, and the second place in English and Portuguese.

Language	EMR	micro P	micro R	micro F1	Accurracy	Rank
English	0.38300	0.46150	0.43020	0.44530	0.88940	2
Portuguese	0.49160	0.55370	0.52630	0.53970	0.90570	2
Russian	0.44390	0.49780	0.48900	0.49330	0.78040	6
Bulgarian	0.51610	0.53970	0.53120	0.53540	0.92740	1
Hindi	0.26900	0.35330	0.29320	0.32050	0.69620	11

Table 2: Official SemEval results on the test set for Subtask 1

4.4 Ablation Experiment

4.4.1 Trial 1: Selecting a model for each language

A multi-turn dialogue training set as shown in A.2 is constructed to train Qwen2.5-32B and Phi-4 with the news articles in each language. In the first round, LLMs are instructed to infer the primary roles of entities based on the original news text. In the second round, models are guided to select fine-grained roles that are consistent with the predicted primary roles. Finally, the results are evaluated using EMR.

The purpose of this experiment is to select the most suitable model for each language. The experimental results are shown in Table 3.

Dataset	Language	Phi-4 EMR%	Qwen2.5 EMR%
EN	English	37.68	**38.06**
PT	Portuguese	42.60	**42.98**
RU	Russian	**40.86**	40.80
BG	Bulgarian	**30.90**	20.36
HI	Hindi	**24.22**	16.76

Table 3: Validation results for Phi-4 and Qwen2.5-32B in five languages

Results and analysis: Phi-4 and Qwen2.5-32B show similar performance in English and Portuguese, with Qwen2.5-32B being slightly better. Phi-4 is significantly more effective in Hindi and Bulgarian, which can be attributed to its extensive support for low-resource languages. Therefore, Qwen2.5-32B is used for English and Portuguese tasks and Phi-4 for Hindi, Bulgarian, and Russian tasks subsequently.

However, neither model shows competitive performance on the training set built from the full texts of the news articles, which shows the necessity of the data processing step in our two-stage news analytical system.

4.4.2 Trial 2: Training with Semantic Segmented Data

As shown in Figure 2, we observe that the same entity can appear multiple times at different positions in an article, each time having a different role. Models struggle to determine which specific entity and location to analyze given the entire news text. To address this issue, we design a target-oriented semantic segmentation method to extract relevant context from news texts according to a specific entity and its specific position. The method is described in Appendix A.1 in detail.

Figure 2: An example of redundant and noisy information for role classification in one article: red represents the *antagonist*, blue represents *protagonist*, green represents *innocent*, and orange represents entities that do not need to be classified. The fine-grained roles of the entities are indicated within the boxes.

The training set is constructed with the same prompt

in trial 1 but replacing full texts for coherent context obtained from semantic segmentation. The results in Table 4 demonstrates the effectiveness of this data processing technique in reducing redundant information and enhancing model accuracy.

Dataset	Language	Phi-4 EMR%	Qwen2.5 EMR%
EN	English	50.46	**50.86**
PT	Portuguese	71.90	**72.80**
RU	Russian	**57.42**	57.06
BG	Bulgarian	**50.73**	43.00
HI	Hindi	**43.57**	39.62

Table 4: Validation results for training with semantic segmented context in five languages

Results and analysis: Both models achieve significant improvements in five languages. The improvement can be attributed to gathering contextual information for entities at specified positions in the original news text, which effectively eliminates redundant information and noise.

4.4.3 Trial 3: Training with Multilingual Data

We observed that data is limited for each language, and a lack of data diversity can lead to suboptimal model performance. Aside from various prompts for rewriting, summarizing, and semantic segmentation, we investigate the use of multilingual datasets as a data augmentation technique to improve model performance and generalization capabilities.

Dataset	Language	EMR%
EN	English	53.35
EN+PT	English	**53.65**
EN+RU	English	52.75
EN+RU+BG	English	45.05
All	English	49.52
PT	Portuguese	75.72
EN+PT	Portuguese	**76.99**
All	Portuguese	74.50
RU	Russian	58.64
EN+RU	Russian	62.79
RU+BG	Russian	58.14
EN+RU+BG	Russian	**65.12**
All	Russian	61.63
BG	Bulgarian	56.86
RU+BG	Bulgarian	**61.29**
EN+RU+BG	Bulgarian	58.14
All	Bulgarian	51.61
HI	Hindi	**48.93**
All	Hindi	48.57

Table 5: Phi-4 on monolingual and multilingual data based on LoRA fine-tuning after 5 models voted after inference on the validation set.

Results and analysis: As shown in Table 5, a combination of the Portuguese and English training sets outperforms both single-language and other multilingual training sets on the validation sets for English and Portuguese. We deduce that both English and Portuguese belong to the Indo-European language family, sharing significant similarities that help the model better capture key information.

For the Bulgarian validation set, a combination of Russian and Bulgarian training sets outperforms other combinations. This might also be attributed to the linguistic similarity between Bulgarian and Russian, which enables knowledge transfer and improves model performance in Bulgarian. However, for the Russian validation set, the combination of English, Russian, and Bulgarian datasets achieves the best performance. The addition of the English dataset enhances the performance in Russian significantly. The reason might require further analysis on data distributions.

For the Hindi validation set, training with Hindi data alone yields better results compared to other multilingual combinations. This is probably because Hindi has a grammatical structure that is significantly different from other languages, and mixing data from other languages does little help in model performance.

This analysis provides some insights into how language similarity influences model performance in multilingual tasks. Despite a minor inconsistency, the takeaway is an actionable guideline for future experiments.

5 System Description for Subtask 2

5.1 Model Design

In Subtask 2, we adopt a similar pipeline for narrative-based semantic segmentation as in Subtask 1. For each news article, we obtain a list of relevant paragraphs for each subnarrative in its golden label set.

Two types of models for multilabel classification are trained, one-vs-rest classification models (referred to as OVR) and label sequence generation models (referred to as SGM). Therefore, two data aggregation approaches are employed to convert the above data into proper training data for different types of models, respectively.

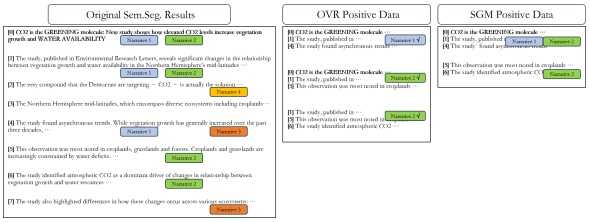

Figure 3: Construct positive training samples for OVR models and SGMs

As demonstrated in the Figure 3, to train classification models, we concatenate several (two to three) relevant paragraphs for each subnarrative to form a large paragraph. Positive data points are made up of these large

paragraphs and their corresponding subnarratives, and negative data points are sampled from combinations of large paragraphs and subnarratives that are not relevant to the entire news article. LLM-based OVR models for multilevel label predictions are trained on multitask datasets that contain training data for coarse and fine labels.

For generation models, a label powerset for each paragraph is first obtained from the original processed data. Then we concatenate several (two to three) paragraphs sharing the same label powerset to form a large paragraph. The volume of data decreases rapidly after concatenation. Therefore, we use various prompts for rewriting, summarizing, and semantic segmentation to expand the training set to contain 10k data points per language (5k for Russian). We train LLM-based sequence generation models for multilevel label predictions using multiturn dialogues that predict coarse and fine labels in turn.

The considerations and effectiveness of the underlined data processing decisions, along with semantic segmentation and OVR and SGM models, will be discussed in the ablation analysis.

5.2 Experimental Setup

We fine-tune the Phi-4 and Qwen2.5 models with LoRA applied to all modules, setting rank = 8, alpha = 16 and dropout of 0.1. We implement distributed training with Deepspeed Zero-2 or Zero-3 on two NVIDIA A100 GPUs with 2 or 4 samples per device and various gradient accumulation parameters to achieve a total batch size of 64. We train the models for 3 epochs and evaluate their performances on the Subtask 2 validation set to select the best checkpoints.

During the prediction phrase, we first perform semantic segmentation without instruction of narratives to divide the entire article into chuncks. Then OVR models are applied to decide whether a chunk is relevant to each narrative and the corresponding subnarratives. SGMs are used to generate a subset of narratives and subnarratives for each chunk. To get the final result for the entire article, we take the union set of the results for all chunks. Finally, we evaluate the results on the validation set using F1 metrics (Piskorski et al., 2025).

5.3 Best Results

The best results on the validation datasets for different languages are obtained by fine-tuned Phi-4 and Qwen2.5-32B-Instruct respectively. Combining English and Portuguese datasets, English, Russian, and Bulgarian datasets leads to further performance gain, which is consistent with the result of Subtask 1. Both multitask OVR models and multiturn SGM models are effective in multilabel classification tasks.

We use the best-performing models for test set inference and the final leaderboard results are presented in Table.

Language	Dataset	Base	Method	F1 samples
English	EN+PT	Qwen2.5-32B	SGM	0.520
Portuguese	EN+PT	Qwen2.5-32B	SGM	0.512
Russian	BG+RU	Phi-4	SGM	0.496
Bulgarian	EN+BG+RU	Phi-4	OVR	0.543
Hindi	HI	Phi-4	OVR	0.457

Table 6: Best results on the validation set for Subtask 2

Language	F1 coarse	F1.st.dev coarse	F1 samples	F1.st.dev samples	Rank
English	0.52100	0.35600	0.33900	0.29100	7
Portuguese	0.54100	0.29000	0.40900	0.26900	2
Russian	0.56600	0.26800	0.43400	0.24700	2
Bulgarian	0.63100	0.33800	0.46000	0.33300	1
Hindi	0.39200	0.39000	0.21800	0.35000	8

Table 7: Official SemEval results on the test set for Subtask 2

5.4 Ablation Experiments

5.4.1 Trial 1: Benefits of Semantic Segmentation

The narrative-oriented semantic segmentation pipeline is critical to enhancing model performance. The necessity of isolating coherent narrative units from an news article is to reduce noisy text spans that are related to other narratives, enabling the model to learn accurate features for each narrative. As shown in Table 8, the models trained on semantically segmented data significantly outperform those trained on full texts by +10pt.

Language	Dataset	Full Text F1 samples	Sem.Seg. F1 samples
English	EN+PT	0.314	**0.432**
Portuguese	EN+PT	0.303	**0.435**
Russian	RU+BG	0.332	**0.419**
Bulgarian	RU+BG	0.354	**0.463**
Hindi	HI	0.279	**0.394**

Table 8: Validation results for training with full texts and semantically segmented data

5.4.2 Trial 2: Data Quality Improvement and Augmentation

We observe that model performance may depend heavily on the quality of the subnarrative labels generated at the paragraph level. Concatenating paragraphs linked to a specific subnarrative or a specific set of subnarratives offers various advantages.

- Balance the specificity and richness of text spans: contain unique utterance that align with a narrative label and their contextual information as well
- Improve the accuracy of positive data points: the accuracy of original paragraph-level labels is around 80% according to manual inspection. By connecting two paragraphs sharing the same label, the accuracy of narrative labels for large paragraphs is around 96% in theory, and over 99% when connecting three. This is of critical importance for sequence generation models because they

do not leverage negative samples during the SFT stage.

For classification models, particular attention should also be paid to the construction of negative samples. It is risky to use narratives that appear in a new article but are not assigned to a paragraph as 'hard negatives', because there is a chance of 10-20% that the sentiment segmentation pipeline failed to recall the paragraph as relevant. Therefore, we construct negative samples with 'easy negatives' and achieve competitive results.

Original news articles are expanded via prompting LLMs for rewriting, summarization, and translation to generate diverse paraphrases while preserving label semantics. We perform semantic segmentation on the generated articles as well to address the rapid decline in training volume after data quality improvement and mitigate overfitting in low-resource scenarios. Data quality improvement and data augmentation yield consistent gains of 5-7pt in F1 samples, as shown in Table 9.

Language	Dataset	Sem.Seg. F1 samples	Sem.Seg.Aug. F1 samples
English	EN+PT	0.432	**0.501**
Portuguese	EN+PT	0.435	**0.483**
Russian	BG+RU	0.419	**0.481**
Bulgarian	BG+RU	0.463	**0.521**
Hindi	HI	0.394	**0.457**

Table 9: Validation results for training with Data Augmentation and Quality Improvement

5.4.3 Trial 3: Comparing OVR and SGM Models

Fine-tuning sequence generation models (SGMs) with multiturn dialogue data further improves task performance. This modeling choice aligns with the hierarchical and interdependent nature of narratives, where coarse labels inform fine-grained labels, and one narrative often co-occurs with a handful of others. The ablation of OVR models and SGMs architectures highlights the efficacy of modeling label dependencies. As shown in Table 10, SGMs outperform OVR models in English, Portuguese, and Russian by 2-3 pt and give comparative results in Bulgarian and Hindi. One limitation is that SGMs only learn from positive labels during SFT. We will explore incorporating negative samples to further improve the performance of SGMs with preference-based reinforcement learning in future experiments.

6 System Description for Subtask 3

6.1 Model Design

Subtask 3 is a task of generating explanations for a given dominant narrative. Despite the outstanding ability of open-source base LLMs for synthetic data generation, there are cases where the models performed poorly in professional areas. Therefore, we employ supervised fine-tuning and reinforcement learning methods to train LLMs and enhance their performance in news analysis.

Language	Dataset	OVR F1 samples	SGM F1 samples
English	EN+PT	0.501	**0.520**
Portuguese	EN+PT	0.483	**0.512**
Russian	BG+RU	0.481	**0.496**
Bulgarian	BG+RU	**0.521**	0.519
Hindi	HI	**0.457**	0.443

Table 10: Validation results for OVR models and SGMs on multilabel classification

To overcome the challenge of limited data, we applied data augmentation through large pre-trained models (Qwen2.5-72B and LLaMa3.3-70B) for text rewriting. We also used resampling techniques to balance the dataset across different languages, ensuring that the model was trained on a representative distribution of data.

For data augmentation, each text sample x_i in the training set was augmented by rewriting it at varying lengths $L \in \{300, 500, 800\}$. The augmented dataset D_{aug} was then formed by mixing the original data and rewritten samples, denoted as:

$$D_{\text{aug}} = \{(x_1, \hat{x}_1^{300}, \hat{x}_1^{500}, \hat{x}_1^{800}), \dots\} \quad (4)$$

where $\hat{x}_i$ represents the augmented version of the original sample x_i.

Reinforcement learning is another technique to improve model performance. The preference data for RL is constructed by leveraging LLMs fine-tuned with SFT datasets. We use fine-tuned models to generate narrative explanations for all data multiple times. The generated explanations for each data point are ranked by their BertScore (Zhang et al., 2019) similarity to the golden label. The top 1 output with a BertScore less than 0.7 is selected as the rejected sequence and the preferred sequence remains the golden label.

Using above methods, we expand the training datasets and empower base LLMs with the ability of generating high-quality outputs in a professional area. Finally, the system achieves outstanding performance in five languages.

6.2 Experiment Setup

we present the design and implementation of the system used for training and fine-tuning multilingual models. The main objective of this work was to explore effective strategies for improving model performance in cases where training data is scarce, leveraging a combination of Supervised Fine-Tuning (SFT), Data Augmentation, and Reinforcement Learning (RL)techniques. The base model we utilized is phi4 (small), with Qwen2.5-72B and LLaMa3.3-70B(AI@Meta, 2024) as the large models for synthetic data generation. Throughout the experiments, we used BertScore as the evaluation metric.

The training process was designed to handle the challenge of limited data for each individual language. To

mitigate this, a multilingual dataset was created by combining five different languages, making the model more adaptable and better at handling diverse linguistic inputs.

6.3 Best Results

6.3.1 Quantitative Analysis

The combination of data augmentation, SFT, and RL produces the best performance on the validation set as shown in Table 11.

Data Augmentation increased the diversity of the training set, enhancing the model's generalization capability and allowing it to perform well across a broader range of inputs. By providing additional data, the model was exposed to various contexts and language patterns, which mitigated the risks of overfitting to a small, homogeneous dataset.

Supervised Fine-Tuning ensured that the models were tailored to the specific task at hand. By fine-tuning the model on a mixed dataset, we enabled it to specialize in producing task-relevant text while still benefiting from a broad linguistic context. This balancing act between generalization and task-specific adaptation is critical for achieving optimal performance.

Preference-Based Reinforcement Learning provided a mechanism for continuous improvement. By incorporating a feedback loop that ranked outputs based on quality metrics like BertScore, the model learned to prioritize high-quality responses. This reinforcement learning step allowed the model to focus not just on producing text, but on generating outputs that were aligned with human preferences and task-specific criteria.

Language	Method	F1 macro
English	DPO	0.75109
Portuguese	DPO	0.75471
Russian	DPO	0.72848
Bulgarian	DPO	0.72466
Hindi	DPO	0.75159

Table 11: Best results on the validation set for Subtask 3

Finally, we achieve first place in Russian, Hindi and Bulgarian, and remarkable results in Portuguese and English on the official SemEval test set leaderboard, as shown in Table 12.

Language	Precision	Recall	F1 macro	Rank
English	0.72371	0.72589	0.72433	8
Portuguese	0.75365	0.73984	0.74637	2
Russian	0.69984	0.71423	0.70639	1
Bulgarian	0.71405	0.69478	0.70396	1
Hindi	0.75097	0.76045	0.75540	1

Table 12: Official SemEval results on the test set for Subtask 3

6.4 Ablation Experiment

This section describes the various experiments conducted to train and fine-tune the models, exploring the impact of different approaches on the model's performance.

6.4.1 Trial 1: Direct Inference with Base Models

Objective: The first experiment aimed to establish a baseline by directly using the base models (phi4) for inference on the validation set without any fine-tuning. The prompt is lited in the appendix B.1.

Methodology: No training was performed, and the raw outputs of the models were evaluated using BertScore to measure the quality of the generated text.

Language	Precision	Recall	F1 macro
English	0.67740	0.72317	0.69948
Portuguese	0.68768	0.74247	0.71378
Russian	0.65268	0.71828	0.68379
Bulgarian	0.65624	0.73164	0.69156
Hindi	0.72861	0.71984	0.72390

Table 13: Results of phi-4 in Direct Inference.

6.4.2 Trial 2: Multilingual SFT Training with Mixed Dataset

Objective: Given the insufficient amount of data for each language, the next experiment involved mixing the five languages into a single training set. This mixed-language dataset was used to train the models.

Methodology: The five languages were combined into one dataset, ensuring that the models had a broader context to learn from. This trial aimed to assess how well the models could handle multiple languages simultaneously.

Language	Precision	Recall	F1 macro
English	0.75589	0.73780	0.74648
Portuguese	0.74036	0.74964	0.74464
Russian	0.73371	0.71030	0.72084
Bulgarian	0.70630	0.71394	0.70963
Hindi	0.75723	0.71460	0.73488

Table 14: Results of phi-4 in Multilingual SFT Training with Mixed Dataset.

Results and analysis: By comparing Table 13 and Table 14, it was found that multilingual mixed SFT fine - tuning is effective, and the model is able to maintain competitive performance across all languages. This experiment highlights the feasibility of training multilingual models with a mixed - language dataset of five languages, indicating that the model can successfully learn from data in different languages.

6.4.3 Trial 3: Data Augmentation through Prompting

Objective: To combat the limited size of the training data, we used Qwen2.5-72B and LLaMa3.3-70B to augment the data. The goal was to generate additional samples that could enhance the model's generalization ability.

Methodology: We prompted the large models to rewrite the data at varying lengths (300, 500, 800 tokens). These augmented data points were mixed with the original training data to create a richer, more diverse training set.

- **Augmented Training Set:** Mixing original and augmented data led to improvements in model performance, especially in cases with a low number of original data points.
- **Balanced Dataset:** To address discrepancies in the amount of data for each language, we employed a resampling technique to balance the data, ensuring that each language had an equal representation in the training set.

Language	Precision	Recall	F1 macro
English	0.76814	0.76814	0.74961
Portuguese	0.73819	0.75801	0.74752
Russian	0.72719	0.72018	0.72306
Bulgarian	0.71685	0.73365	0.71704
Hindi	0.74930	0.74930	0.74368

Table 15: Results of phi-4 in Multilingual SFT Training with Data Augmentation and Random Replication.

Results and analysis: A comparison of Table 14 and Table 15 reveals that model performance can be further enhanced through data augmentation. This is attributed to the use of large models such as Qwen2.5 - 72B and LLaMa3.3 - 70B to generate additional data samples of varying lengths, as well as the resampling method employed to balance the representation of each language. By enriching the training set in this way, we have strengthened the model's generalization ability, thus demonstrating the effectiveness of data augmentation techniques in improving model performance.

6.4.4 Trial 4: Preference-Based Reinforcement Learning

Objective: In this trial, we aimed to improve the quality of the generated data by constructing preference data, and then fine-tuning the models using Reinforcement Learning based on these preferences.

Methodology: We first generated outputs for both the original and augmented datasets using the model weights from the previous experiment. These outputs were scored using BertScore, and the highest-scoring samples (top 2-3) were selected as negative samples for the RL training.

Results and analysis: As shown in Table 16, the model has achieved good performance across various languages, with precision, recall, and F1-score consistently remaining above 0.7, which is a further improvement compared to SFT fine-tuning. This is attributed to our use of BertScore to rank the outputs of the original dataset and the augmented dataset. We selected the top - scoring samples as negative samples for RL training. This preference-based RL fine-tuning can assist the model in generating more relevant and higher-quality responses, indicating that our approach is effective.

Language	Precision	Recall	F1 macro
English	0.76102	0.74214	0.75109
Portuguese	0.75258	0.75752	0.75471
Russian	0.74098	0.71728	0.72848
Bulgarian	0.71685	0.73365	0.72466
Hindi	0.76290	0.74152	0.75159

Table 16: Results of phi-4 with SFT and DPO Negative Samples.

7 Conclusion

For Subtask 1 and 2 in SemEval-2025 Task 10, we perform semantic segmentation through narrative- or entity-based prompt engineering, obtaining the most relevant contextual information for each narrative or entity to reduce redundant and interfering information in classification. We also analyze the effectiveness of monolingual and multilingual training approaches in Subtask 1 and 2, and observe that multilingual datasets composed of adjacent languages achieve better results than monolingual datasets. In Subtask 3, we applied data augmentation and achieved strong results using SFT fine-tuning combined with preference-optimized reinforcement learning.

Our systems incorporate various state-of-the-art techniques, including prompt-engineering, LoRA fine-tuning, reinforcement learning, and LLM for data augmentation, and achieve a total of five first places and five second places in five languages. The results demonstrate the effectiveness of the proposed methods in entity framing, narrative classification, and narrative extraction from news articles.

To improve our work in the future, we plan to look deeper into the lexical or label distribution differences between the training set and the test set to help build more robust systems.

8 Limitation

In order to understand the errors made by the LLMs, we conduct a manual review of the generated content. While the system performed well in generating responses, some common error types were observed:

- Syntax Errors: LLMs sometimes struggle with complex sentence structures, particularly in lan-

guages with more intricate syntactical rules such as Hindi and Russian.

- Lexical Errors: LLMs occasionally selected words that were contextually inappropriate due to lack of domain knowledge. This was more apparent in languages with less raining corpora (e.g., Hindi).
- Translation Errors: In translation for data augmentation, LLMs sometimes failed to translate words or sentence structures accurately, especially between languages with significant morphological differences (e.g., Hindi and Russian).
- Overfitting to Augmented Data: After finetuning with augmented data, LLMs sometimes generate repetitive and verbose responses, which may be signs of overfitting to augmented data.

This analysis helps understand the limitations of LLMs in news analysis and areas that reuqire future refinement for complex sentence understanding and generation tasks.

Acknowledgments

This task has been accomplished with the funding of *Ping An Life Insurance Company of China, Ltd*. All research presented in this paper was conducted during the Semeval-2025 competition. The opinions and conclusions expressed herein are solely those of the authors. We would also like to thank the task organizers and anonymous reviewers for their valuable feedback and constructive comments.

References

Marah Abdin, Jyoti Aneja, Harkirat Singh Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C'esar Teodoro Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *ArXiv*, abs/2412.08905.

AI@Meta. 2024. Llama 3 model card.

Alberto Barrón-Cedeño, Firoj Alam, Tanmoy Chakraborty, Tamer Elsayed, Preslav Nakov, Piotr Przybyła, Julia Maria Struß, Fatima Haouari, Maram Hasanain, Federico Ruggeri, Xingyi Song, and Reem Suwaileh. 2024. The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In *Advances in Information Retrieval*, pages 449–458, Cham. Springer Nature Switzerland.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. *CoRR*, abs/2105.09284.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. *arXiv preprint arXiv:1805.11004*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificaccao Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, Jakub Piskorski, and Preslav Nakov. 2025. Entity framing and role portrayal in the news.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. 2023. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, European Commission Joint Research Centre, Ispra (Italy).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Dymitr Ruta and Bogdan Gabrys. 2005. Classifier selection for majority voting. *Information fusion*, 6(1):63–81.

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

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Xue Wan, Wensheng Zhang, Mengxing Huang, Siling Feng, and Yuanyuan Wu. 2023. A unified approach to nested and non-nested slots for spoken language understanding. *Electronics*, 12(7):1748.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at SemEval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 488–497, Toronto, Canada. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yong Zhang, Hongrui Zhang, Jing Cai, and Binbin Yang. 2014. A weighted voting classifier based on differential evolution. In *Abstract and applied analysis*, volume 2014, page 376950. Wiley Online Library.

Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical metadata-aware document categorization under weak supervision. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 770–778.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Fan Zhou, Zengzhi Wang, Qian Liu, Junlong Li, and Pengfei Liu. 2024. Programming every example: Lifting pre-training data quality like experts at scale. *arXiv preprint arXiv:2409.17115*.

A Subtask 1

A.1 Article Preprocessing

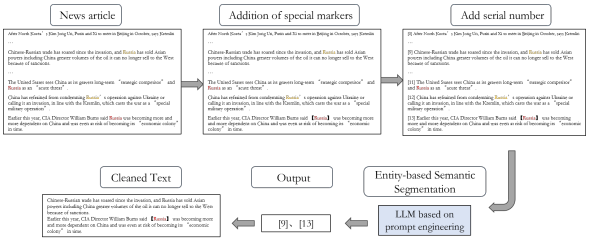

Figure 4: An example of a data preprocessing workflow for context aggregation based on entity positions in news text.

First, we annotate the entities to be classified in the news text using a special marker. Then, we slice the news text by newline characters and sequentially add a number at the beginning of each segment (e.g., [1] Text.), followed by concatenating the segments in order using newline characters. Next, the processed news text is input into the Qwen2.5-72B model, based on prompt engineering, to output the corresponding context segment numbers for each entity. Finally, based on the sequence of output numbers, we concatenate the cleaned segments using newline characters to obtain the text with redundant information and noise removed.

A.2 Fine-tuning Prompt

In Figure 5, the numbers highlighted in red indicate the corresponding input information: Number ① represents the input belonging to the Ukraine-Russia War or Climate Change domain; Number ② represents the entity to be predicted; Number ③ represents the input news text; Number ④ represents the output of the first-round dialogue (prediction or label) for the primary role; Number ⑤ represents the input for the entity to be predicted. By constructing the two-round dialogue fine-tuning prompt in this way, the model can focus on the relationship between the primary role and the fine-grained roles, thereby improving its performance.

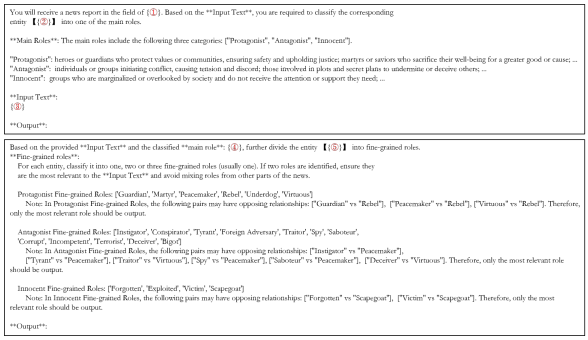

Figure 5: An example of prompt engineering fine-tuning based on a multi-turn dialogue. In the first round (shown above), the prompt engineering is used to guide the model in predicting the main role of the corresponding entity. In the second round (shown below), the prompt engineering utilizes the main role predicted in the first round to predict the fine-grained roles.

B Subtask 3

B.1 Fine-tuning Prompt

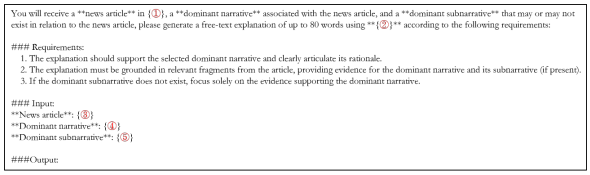

Figure 6: A example of a prompt for fine-tuning LLM based on LoRA.

In Figure 6, the numbers highlighted in red indicate the required input information: Number ① represents the domain of the input, either Ukraine-Russia War or Climate Change; Number ② represents the language in which the output is provided; Number ③ represents the input news text; Number ④ represents the main narrative; and Number⑤ represents the secondary narrative.