BERTastic at SemEval-2025 Task 10: State-of-the-Art Accuracy in Coarse-Grained Entity Framing for Hindi News

Tarek Mahmoud, Zhuohan Xie, Preslav Nakov
Mohamed Bin Zayed University of Artificial Intelligence
{tarek.mahmoud,preslav.nakov}@mbzuai.ac.ae

Abstract

We present our system and share insights for SemEval-2025 Task 10 Subtask 1 on coarse-grained entity framing in Hindi news. Our work investigates two complementary strategies. First, we explore LLM prompting with GPT-4o, comparing multi-step hierarchical prompting against naïve single-step prompting for both main and fine-grained role prediction. Second, we conduct an extensive study on fine-tuning XLM-R, evaluating performance across different context granularities (i.e., using the full text of the news article, or the paragraph or the sentence containing the entity mention) and contrasting monolingual versus multilingual settings, as well as training on main role versus fine-grained role labels. Among all configurations, the best system was trained on fine-grained role annotations on all languages using the sentence context, and it achieved an exact match ratio of 43.99%, micro precision of 56.56%, micro recall of 47.38%, and a micro F1 of 51.57%. Notably, our system attained a state-of-the-art main role accuracy of **78.48%** on Hindi news—surpassing the next best result of 76.90%—as demonstrated on the official test leaderboard.[1] Our findings offer valuable insights into effective strategies for robust entity framing in multilingual and low-resource settings.

1 Introduction

Multilingual news narrative analysis has become increasingly important in the digital age. The rapid proliferation of social media has transformed information dissemination, enabling instant news access and the global spread of narratives. However, this connectivity also amplifies risks such as biased reporting, propaganda, and narrative manipulation. These are challenges that are particularly evident during conflicts and political upheavals. SemEval-2025 Task 10 on Multilingual Characterization and Extraction of Narratives from Online News addresses these challenges by providing a platform to study entity framing across five languages (Bulgarian, English, Hindi, European Portuguese, and Russian) (Piskorski et al., 2025). Our work specifically focuses on coarse-grained entity framing in Hindi news, an important subset given its potential impact on public perception.

Entity framing is an interesting task that transcends surface-level lexical features to uncover the subtle semantic nuances in how entities are portrayed. It focuses on interpreting the contextual cues that reveal whether the entity is depicted as a protagonist, antagonist, or an innocent bystander. This analysis is fundamental for understanding media bias, as it illuminates how news authors strategically frame narratives to influence public perception and shape discourse. The entity framing task is described in-depth in Mahmoud et al. (2025); Stefanovitch et al. (2025).

Our experiments fall under two main approaches.

1. First, we explore LLM prompting with GPT-4o (OpenAI, 2024), testing a range of prompting techniques—including single-step, and multi-step, hierarchical methods to predict both main and fine-grained role labels.

2. Second, we fine-tune XLM-R (Conneau et al., 2020) to examine the effect of different context granularities (for example, using the sentence versus the paragraph containing the entity mention, or the full text) and to compare monolingual versus multilingual training setups, as well as training on main role versus fine-grained role labels.

This comprehensive investigation enables us to determine the optimal configuration for capturing

[1] The leaderboard is available at https://propaganda.math.unipd.it/semeval2025task10/leaderboard.html

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 386–396
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

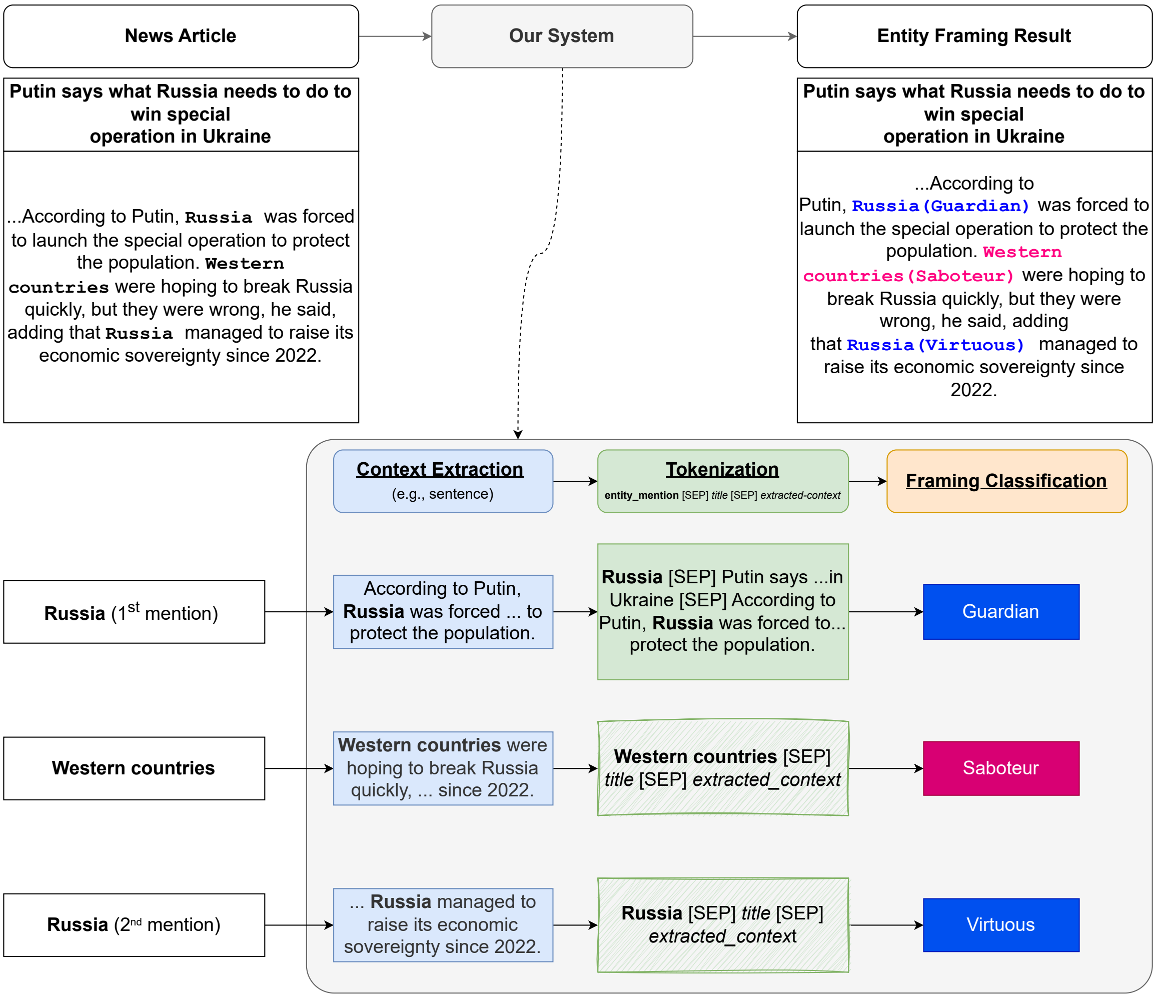
News Article
Our System
Entity Framing Result
Putin says what Russia needs to do to win special operation in Ukraine
...According to Putin, Russia was forced to launch the special operation to protect the population. Western countries were hoping to break Russia quickly, but they were wrong, he said, adding that Russia managed to raise its economic sovereignty since 2022.
Putin says what Russia needs to do to win special operation in Ukraine
...According to Putin, Russia (Guardian) was forced to launch the special operation to protect the population. Western countries (Saboteur) were hoping to break Russia quickly, but they were wrong, he said, adding that Russia (Virtuous) managed to raise its economic sovereignty since 2022.
Context Extraction
(e.g., sentence)
Tokenization
entity_mention [SEP] title [SEP] extracted-context
Framing Classification
Russia (1st mention)
According to Putin, Russia was forced ... to protect the population.
Russia [SEP] Putin says ...in Ukraine [SEP] According to Putin, Russia was forced to... protect the population.
Guardian
Western countries
Western countries were hoping to break Russia quickly, ... since 2022.
Western countries [SEP] title [SEP] extracted_context
Saboteur
Russia (2nd mention)
... Russia managed to raise its economic sovereignty since 2022.
Russia [SEP] title [SEP] extracted_context
Virtuous

Figure 1: **A simplified overview of our approach:** The figure illustrates how we approach the entity framing task. Given an article, and a list of entity mentions (shown in boldface) for which framing is needed, we process one entity mention at a time. We then extract the context surrounding the entity mention (e.g., the sentence containing the entity mention). We tokenize the input if necessary, then we obtain the framing classification for that entity mention from the model (XLM-R or GPT-4o). This process is repeated for each entity mention.

the intricate nuances of entity framing by balancing zero-shot prompting with fine-tuning strategies using state-of-the-art LLMs. In addition, our analysis provides valuable insights into the unique strengths and weaknesses of each approach, informing future research directions in multilingual narrative analysis.

On the official test set leaderboard, our best system—trained on all languages using fine-grained role annotations with sentence-level context—achieves an exact match ratio of 43.99%, micro precision of 56.56%, micro recall of 47.38%, and a micro F1 of 51.57%, and, notably, our system reached a state-of-the-art main role accuracy of **78.48%**, ranking the top on main role accuracy on the leaderboard on Hindi news. While these results illustrate the effectiveness of our approach, they also highlight ongoing challenges in modeling context and subtle linguistic cues in multilingual settings.

2 Background

2.1 Task Formulation

Entity framing is the task of assigning one or more roles to each named entity mention in a news article according to a two-level hierarchical taxonomy that has a rich set of fine-roles exceeding twenty roles all nested under three main roles: protagonist, antagonist, and innocent. Given an article and a list of entity mentions (each specified by its text span and offsets), the goal is to predict a coarse main role—selected from *protagonist*, *antagonist*, or *innocent*, as well as one or more fine-grained sub-roles. For example, an entity such as "NATO" might be labeled as an *protagonist* with sub-roles like "guardian" and "virtuous". This formulation is cast as a multi-label, multi-class text-span classification problem and is crucial for analyzing how news narratives construct public perception. Additionally, when focusing solely on the main role prediction, the problem simplifies to a multi-class classification task at the coarse-grained level of the taxonomy.

The shared task dataset consists of 1,378 recent news articles in five languages (Bulgarian, English, Hindi, European Portuguese, and Russian) covering two globally significant domains: the Ukraine-Russia conflict and climate change. In total, the dataset has over 5,800 entity mentions that have been annotated with detailed role labels according to a hierarchical taxonomy developed by the shared task organizers.

2.2 Related Work

We explore two main approaches for modeling entity framing:

For our first approach, we fine-tune XLM-R and investigate various context granularities (sentence, paragraph, full text) surrounding the entity mention. We also compare monolingual and multilingual performance. Entity framing closely resembles targeted sentiment analysis (or Aspect-Based Sentiment Analysis, ABSA). For instance, Bastan et al. (2020) compare document- and paragraph-level contexts for target-based sentiment analysis, and their results support our intuition that a narrower context improves accuracy. However, their work, focused on English news, is less complex than our multilingual hierarchical entity framing task, and they do not explore sentence-level granularity. In addition, Goyal et al. (2021); Downey et al. (2024) and Chen et al. (2024) show that performance improves as more languages are included during training, motivating our multilingual experiments.

For our second approach, we evaluate naïve prompting against a variant of least-to-most prompting (Zhou et al., 2023), which we refer to as hierarchical prompting. Note that our hierarchical prompting differs from the reasoning frameworks of Sridhar et al. (2023) and Budagam et al. (2024). We name our method hierarchical prompting because the taxonomy is hierarchical; we break down the entity framing task into two stages: first, we identify the main role, and second, we determine the underlying fine-grained role.

3 System Overview

In this section, we describe our experimental methods for framing classification, outlining critical modeling decisions. We present two core strategies: fine-tuning multilingual Transformers and applying hierarchical prompting with LLMs. Key details of our approach are shown in Figure 1.

3.1 Fine-Tuning Pretrained Multilingual Transformers

We fine-tune XLM-R, a multilingual Transformer model, to classify entities at both the **main role level** (protagonist, antagonist, or innocent) and the **fine-grained role level** (22 subcategories). Our design choices focused on:

- **Granularity of Context**: We experiment with different context windows—full document, paragraph, and sentence-level—to assess the impact of surrounding text on classification accuracy.

- **Multilingual vs. Monolingual Training**: We compare the performance of models trained on a single language with models trained on data from all five languages.

- **Handling Span-Level Classification**: To address span-based classification within long documents, we structure input text as:

  ```
  input = entity mention + [SEP] + title
  + [SEP] + context
  ```

  where context varied depending on the granularity setting.

- **Loss Function and Training Strategy**: We use a softmax activation for multi-class classification at the main role level and a sigmoid activation for multi-label fine-grained role classification. The model was optimized using **Binary Cross-Entropy loss**, allowing it to predict multiple overlapping roles per entity span.

3.2 Hierarchical Zero-Shot Learning with LLMs

To leverage the capabilities of state-of-the-art LLMs, we explore two prompting strategies:

- **Single-Step Prompting**: This is the naïve approach where we predict both the main role and fine-grained role in a single prompt, relying on the model's ability to process the entire task holistically.

- **Hierarchical Multi-Step Prompting**: This approach decomposes the classification task into two stages:
  - **Step 1**: Predict the main role (protagonist, antagonist, or innocent).
  - **Step 2**: Based on the main role prediction, predict the underlying fine-grained role.

This hierarchical strategy follows the "least-to-most" prompting paradigm, allowing the model to break down complex tasks into sequential reasoning steps. For details on the prompts used in both the single- and multi-step approaches, see Section A, and refer to Section C for the experimental settings.

Rank	Team	Main Role Acc.
1	BERTastic (Ours)	**78.48%**
2	QUST	76.90%
3	DEMON	75.63%
4	gowithnlp	71.52%
5	Dhananjaya	70.89%
6	PATeam	69.62%
7	FromProblemImportSolve	66.77%
8	Fane	65.51%
9	LATeIIMAS	63.61%
10	DUTIR	59.81%
11	Cimba	47.15%
12	LTG	37.66%
13	HowardUniversityAI4PC	35.44%
14	TartanTritons	32.91%
15	Baseline	32.28%

Table 1: Official Test Set Performance on Hindi Main Role Accuracy.

4 Results

In this section, we report the performance of our system on the SemEval-2025 Task 10 as well as on additional experiments conducted using the development set. We focus our attention on main role accuracy, while noting that the Exact Match Ratio on the fine-grained roles is the official leaderboard metric.

4.1 Official Test Set Performance on Hindi Main Role Accuracy

For the official test set, we focus our discussion exclusively on the Hindi news track and, in particular, on the main role prediction—i.e., determining whether an entity is framed as a *protagonist*, *antagonist*, or *innocent*. According to the official leaderboard,[2] our system achieved a main role accuracy of **78.48%** on Hindi news articles. This performance surpasses that of the closest competing system, QUST (76.90%) and DEMON (75.63%), and significantly exceeds the main role accuracies reported by other teams. These results, displayed in Table 1, underscore the robustness of our approach in capturing the nuanced framing of entities in Hindi news narratives.

[2] For full leaderboard details, see `https://propaganda.math.unipd.it/semeval2025task10/leaderboard.html`

Method	Lang.	Main Role		Fine Grained Role				Cost (USD)
		Accuracy	Balanced Accuracy	P	R	Micro F1	Macro F1	
Single-Step LLM Prompting	EN	.8346	.6756	.2692	.4632	.3405	.2171	0.7989
	BG	.8065	.7380	.3725	.5588	.4471	.3481	0.2751
	HI	.6327	.6247	.2753	.4000	.3262	.2196	2.4696
	PT	.7812	.7455	.5167	.6643	.5813	.2891	1.0200
	RU	.7558	.6719	.3939	.5843	.4706	.4644	0.7587
	All	.7030	.6957	.3211	.4726	.3824	**.3103**	5.3223
Multi-Step LLM Prompting	EN	.8031	.6799	.2887	.4118	.3394	.2383	0.5130
	BG	.8031	.6799	.4318	.5588	.4872	.3601	0.5130
	HI	.6367	.6284	.2676	.2868	.2769	.1771	1.4581
	PT	.8125	.7882	.3895	.2643	.3149	.2498	0.5634
	RU	.7442	.6680	.4118	.4719	.4398	.3774	0.4769
	All	.7053	.7017	.3051	.3294	.3168	.2765	**3.1852**
XLM-R	EN	0.6889	0.5276	.1854	.2828	.2240	.1327	–
	BG	0.7333	0.5791	.3030	.3030	.3030	.1349	–
	HI	0.7025	0.7046	.3234	.4951	.3912	.2043	–
	PT	0.8957	0.8840	.6259	.7480	.6815	.2040	–
	RU	0.8000	0.7604	.4831	.4886	.4859	.2364	–
	All	**0.7529**	**0.7553**	**.3649**	**.4985**	**.4213**	.2392	–

Table 2: Consolidated results on the **development set** comparing fine-tuning XLM-R and zero-shot learning with GPT-4o. The table shows performance and cost comparisons between single-step and multi-step LLM prompting approaches, where the highest scores between these two approaches across all languages are underlined. The top results across all three methods and languages are highlighted in **bold**.

4.2 Additional Experiments

Before our official submissions, we conducted a comprehensive set of experiments on the development set (using an 80/20 split on the training data) to analyze the impact of context granularity, training configuration, and modeling strategies on both main and fine-grained role classification. We have also reported the same results in Mahmoud et al. (2025). Among all systems, the XLM-R model trained on all languages using sentence-level context performed best on the development set; therefore, our official leaderboard results are based on this system.

4.2.1 Fine-Tuning Multilingual Transformers

Context Granularity We fine-tuned the multilingual Transformer XLM-R for both the main role (3-class) and fine-grained role (22-class) classification tasks. In our experiments, we explored various context granularities by restricting the input to the full document, the paragraph, or the sentence containing the entity mention. Table 3 summarizes our findings:

- **Main Role Classification:** Models trained on main role labels performed best with paragraph-level contexts, followed closely by sentence-level contexts; document-level contexts resulted in the poorest performance.

- **Fine-Grained Role Classification:** When training on fine-grained labels, sentence-level contexts yielded the highest micro and macro F1 scores, with paragraph-level contexts offering comparable yet slightly lower performance.

Multilingual versus Monolingual Training Table 4 provides insights into the performance of XLM-R when fine-tuned on either monolingual data or a combined multilingual dataset. Our results indicate that while monolingual models exhibit varied performance across languages, the multilingual setting consistently outperforms the monolingual approach. Notably, even though the macro F1 scores remain low—reflecting the challenge of predicting rare fine-grained roles—the multilingual model better captures the diverse linguistic patterns across the five target languages.

4.2.2 Hierarchical Zero-Shot Learning

We further evaluated two prompting strategies using GPT-4o: a *single-step* approach that jointly predicts main and fine-grained roles, and a *multi-step* (hierarchical) approach that decouples the prediction into sequential stages. As shown in Table 2, the multi-step approach yielded a modest improvement in main role prediction over the single-step method, likely due to its focused decomposition of the task. However, the single-step approach demonstrated better performance for fine-grained role predictions, suggesting that joint reasoning may better capture dependencies between roles. Additionally, the multi-step approach proved to be more cost-effective in terms of token usage.

4.3 Discussion

While both approaches are complementary, the fine-tuning method is particularly sensitive to data imbalance. Our tokenization strategy results in frequent entity mentions disproportionately influencing the model weights, often overshadowing the rarer mentions. Zero-shot prompting overcomes these limitations by not relying on training data. In our experiments, XLM-R outperforms zero-shot methods on all evaluation metrics except Macro F1, where it records the lowest performance. We hypothesize that this discrepancy arises from the limited training instances for rare roles, which hampers XLM-R's ability to learn these categories effectively. In contrast, zero-shot approaches remain unconstrained by training data volume and thus maintain robustness for infrequent roles.

Collectively, these additional experiments demonstrate that localized context (i.e., paragraph- or sentence-level) and multilingual training significantly enhance entity framing performance. While the main role accuracy improves with focused context, the challenge of predicting rare fine-grained roles persists, as evidenced by the lower macro F1 scores.

In summary, our system not only delivers state-of-the-art performance on the official test set for Hindi main role prediction but also shows robust performance across various experimental settings on the development set. These results validate the effectiveness of our approach—leveraging both fine-tuning and hierarchical zero-shot prompting—to explore the intricate nuances of entity framing in multilingual news narratives.

5 Conclusion and Future Work

In this work, we present a comprehensive investigation into multilingual entity framing for news narrative analysis, with a focus on coarse-grained framing in Hindi news. We consider two complementary approaches: (1) fine-tuning XLM-R under different context granularities, comparing monolingual and multilingual training setups as well as main versus fine-grained role label training, and (2) zero-shot prompting with GPT-4o, where we explore single-step and hierarchical prompting strategies to predict both main and fine-grained role labels

Our experiments demonstrate that leveraging sentence-level context and training on a diverse multilingual corpus yield superior performance, with our system achieving a state-of-the-art main role accuracy of **78.48%** on Hindi news. These results confirm that carefully selecting the appropriate context window is critical to capturing the subtle semantic nuances of entity framing. While both approaches provide distinct advantages—fine-tuning leverages deep contextual representations to better model linguistic cues, and prompting enables rapid adaptation and zero-shot inference—each also faces challenges in processing the inherent complexities of multilingual narrative content.

Looking forward, our findings motivate further research on refining context-aware strategies and exploring more advanced fusion techniques for integrating heterogeneous contextual cues. Future work may also extend our methods to a broader array of languages and narrative domains, ultimately contributing to the development of robust tools for detecting media bias and understanding public perception in a global news landscape.

Acknowledgments

We would like to thank the anonymous reviewers for their time and valuable insights.

References

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Devichand Budagam, Ashutosh Kumar, Mahsa Khoshnoodi, Sankalp KJ, Vinija Jain, and Aman Chadha. 2024. Hierarchical prompting taxonomy: A universal evaluation framework for large language models aligned with human cognitive principles.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. 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*, pages 8440–8451, Online. Association for Computational Linguistics.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. Targeted multilingual adaptation for low-resource language families.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificação 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.

Team OpenAI. 2024. Gpt-4o system card.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo 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.

Abishek Sridhar, Robert Lo, Frank F. Xu, Hao Zhu, and Shuyan Zhou. 2023. Hierarchical prompting assists large language model on web navigation.

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).

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.

Appendix

A Prompts for Hierarchical Zero-Shot Experiments

```
You are an expert at identifying entity framing and role portrayal in news articles. Analyze the following entity
mention in context, and predict its main role and fine-grained role(s) from the taxonomy below.

Taxonomy: {detailed taxonomy with definitions and examples}

Context Around Entity: {context}

Entity Mention: {entity mention}

Task: Based on the provided context, assign to the entity mention at least one fine-grained role and
exactly one main role.

Return a JSON that has below attributes:
- main role: either one of Protagonist, Antagonist, or Innocent
- fine grained roles: a list of all your predicted fine-grained roles
```

Figure 2: **Single-Step Prompt Template.** The detailed taxonomy includes the roles and their detailed definitions as provided by in the shared task. The context is the text consisting of entity mention along with the 20 words before and after the entity mention.

```
First Step (LLM Call 1): Predict the Main Role

You are an expert at identifying entity framing and role portrayal in news articles. Analyze the following entity
mention in context, and predict its main role from the taxonomy below.

Taxonomy: {list of fine-grained roles per main role}

Context Around Entity: {context}

Entity Mention: {entity mention}

Task: Based on the provided context, assign to the entity mention exactly one main role.

Return a JSON that has this attribute:
- main role: either one of Protagonist, Antagonist, or Innocent


Second Step (LLM Call 2): Predict the Fine-Grained Role

You are an expert at identifying entity framing and role portrayal in news articles. This entity is
portrayed as a(n) {main role} and your task is to analyze the entity mention in context
and predict its fine-grained role(s) from the taxonomy below.

Taxonomy: {pertinent portion of the detailed taxonomy with definitions and examples}

Context Around Entity: {context}

Entity Mention: {entity mention}

Task: Based on the provided context, assign to the entity mention at least one fine-grained role.

Return a JSON that has this attribute:
- fine grained roles: a list of all your predicted fine-grained roles
```

Figure 3: **Multi-Step Prompt Template.** In the first step, the taxonomy is only the tree structure of the taxonomy and does not include any definitions or examples. In the second step, the detailed taxonomy only includes the branch under the predicted main role in the first step. The context is as defined in Figure 2.

B Results on Additional Experiments

Train	Context	Main Role		Fine Grained Role			
		Accuracy	Balanced Accuracy	P	R	Micro F1	Macro F1
M	DOC	.6010	.5904	–	–	–	–
	PAR	.7379	.7385	–	–	–	–
	SEN	.7179	.7123	–	–	–	–
F	DOC	.7229	.7235	.3495	.4446	.3913	.2306
	PAR	**.7529**	**.7553**	.3649	.4985	.4213	.2392
	SEN	.7496	.7503	**.4195**	**.4492**	**.4339**	**.2529**

Table 3: Performance of entity framing on the **development set** across different granularity settings using XLM-R trained on the full multilingual dataset. Models are trained and evaluated on texts with varying context sizes: full document (DOC), paragraph (PAR), or sentence (SEN) containing the entity mention. The results cover models trained on main roles (M), fine-grained roles (F), and evaluated on either main roles, fine-grained roles, or both.

(a) Monolingual setting

Lang.	P	R	Micro F1	Macro F1
EN	.1032	.1313	.1156	.0435
BG	.1056	.5758	.1784	.0505
HI	.3424	.4495	.3887	.1740
PT	.6124	.6423	.6270	.1505
RU	.1077	.5227	.1786	.0437

(b) Multilingual setting

Lang.	P	R	Micro F1	Macro F1
All	.3649	.4985	.4213	.2392
EN	.1854	.2828	**.2240**	**.1327**
BG	.3030	.3030	**.3030**	**.1349**
HI	.3234	.4951	**.3912**	**.2043**
PT	.6259	.7480	**.6815**	**.2040**
RU	.4831	.4886	**.4859**	**.2364**

Table 4: Results on the **development set** for multi-label fine-grained role classification with XLM-R trained on monolingual and multilingual data (evaluated at the paragraph level).

C Experimental Settings

All fine-tuning experiments were conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory. We fine-tuned XLM-R (XLM-RoBERTa) in a single run, using a fixed random seed to ensure reproducibility. When the input context was at the sentence granularity, we performed sentence splitting using Stanza pipelines for each one of our five languages. For XLM-R, default settings were applied, with the following configurations:

- Model: XLM-R$_{\text{base}}$ (125M parameters)
- Learning Rate: 2e-5
- Batch Size: 8
- Epochs: 20 (with early stopping of 3 based on validation loss)
- Random Seed: 42
- Weight Decay: 0.01

To optimize performance, the sigmoid thresholds for fine-grained role predictions were tuned on the validation set. These optimized thresholds were then applied to generate predictions on the test set.

To prevent data leakage, we created train/validation/development splits based on entire articles rather than individual entity-mention annotations. The details of these splits are provided in Table 5. Note that we use the development set provided by the shared task as is, but we split the provided training set using an 80/20 split into train and validation sets. In doing so, the development set acts as a test set in a realistic scenario allowing us to create robust models enabling better generalization of our results to the official test set.

	BG	EN	HI	PT	RU	All
Train	165 (389)	133 (440)	203 (1347)	206 (833)	89 (252)	796 (3261)
Validation	94 (237)	69 (245)	139 (983)	100 (417)	44 (114)	446 (1996)
Dev	15 (30)	27 (90)	35 (279)	31 (115)	28 (85)	136 (599)
Total	274 (656)	229 (775)	377 (2609)	337 (1365)	161 (451)	1378 (5856)

Table 5: Distribution of articles and entity mentions by language and split. The number of entity mentions is shown in parentheses

For the zero-shot experiments, we used OpenAI's GPT-4o (gpt-4o-2024-11-20) with a temperature setting of 0.2 to produce more conservative responses. To ensure the outputs conformed to our defined data types, we employed OpenAI's Structured Outputs API, which returned results in the expected JSON format.