TartanTritons at SemEval-2025 Task 10: Multilingual Hierarchical Entity Classification and Narrative Reasoning using Instruct-Tuned LLMs

R Raghav[*1] **Adarsh Prakash Vemali**[*2]
Darpan Aswal[3] **Rahul Ramesh**[1] **Parth Tusham**[4] **Pranaya Rishi**[5]

[1]Independent Researcher [2]University of California, San Diego
[3]Université Paris-Saclay [4]Texas A&M University [5]Irvington High School
{rraghav5600, vemali.adarsh}@gmail.com

Abstract

In today's era of abundant online news, tackling the spread of deceptive content and manipulative narratives has become crucial. This paper details our system for SemEval-2025 Task 10, focusing on Subtasks 1 (Entity Framing) and 3 (Narrative Extraction). We instruct-tuned quantized Microsoft's Phi-4 model, incorporating prompt engineering techniques to enhance performance. Our approach involved experimenting with various LLMs, including LLaMA, Phi-4, RoBERTa, and XLM-R, utilizing both quantized large models and non-quantized small models. To improve accuracy, we employed structured prompts, iterative refinement with retry mechanisms, and integrated label taxonomy information. For subtask 1, we also fine-tuned a RoBERTa classifier to predict main entity roles before classifying the fine-grained roles with Phi-4 for the English language. For subtask 3, we instruct-tuned Phi-4 to generate structured explanations, incorporating details about the article and its dominant narrative. Our system achieves competitive results in Hindi and Russian for Subtask 1.

1 Introduction

The internet has facilitated direct communication between information producers and consumers, making it easier for deceptive content and manipulative narratives to spread. To address this challenge, SemEval-2025 Task 10 (Piskorski et al., 2025) introduces the "Multilingual Characterization and Extraction of Narratives from Online News". The task spans over five languages - Bulgarian, English, Portuguese, Hindi, and Russian. The task has three subtasks - Entity Framing, Narrative Classification and Narrative Extraction.

In subtask 1, the dataset has news articles which have entities that need classification into main roles and then further into corresponding fine-grained roles. Each article may have multiple instances of the entity but the task entails classification on a specific occurence. Similar to Subtask 1, Subtask 2 focuses on assigning a single dominant narrative and one or more associated sub-narratives to a given news article. Subtask 3 involves generating a concise explanation that supports the dominant narrative of a news article, given the article and the narratives.

Language	Subtask 1		Subtask 3	
	Baseline	System	Baseline	System
Bulgarian	0.04	0.41 (5^{th})	0.63	0.67 (4^{th})
English	0.03	0.36 (5^{th})	0.67	0.72 (9^{th})
Portuguese	0.05	0.33 (8^{th})	0.68	0.70 (5^{th})
Hindi	0.06	0.45 (2^{nd})	0.67	0.70 (4^{th})
Russian	0.05	0.47 (3^{rd})	0.64	0.68 (3^{rd})

Table 1: Leaderboard scores and rankings on the test set. Rankings are against 34 and 17 teams in Subtask 1 and 3 respectively across languages

We participated in Subtasks 1 and 3, by instruct-tuning Microsoft's Phi-4 (Abdin et al., 2024) model and various prompt engineering techniques specific to these subtasks. Our approach experimented with multiple Large Language Models (LLMs), specifically LLaMA (Touvron et al., 2023), Phi-4, RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2019). We utilized both quantized large models and non-quantized small models, discovering that the former performed better. To enhance classification accuracy, we structured prompts to enforce output format consistency, used an iterative refinement methodology with retry mechanisms to reduce LLM generation errors, and incorporated label taxonomy information in prompts to improve performance. For Subtask 1, we also fine-tuned a RoBERTa classifier to predict main entity roles (with 72% accuracy) before refining for fine-grained role classification with Phi-4 for the English language. For other languages, prompt engineering with instruction tuning to generate

[*]Equal contribution

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1964–1973

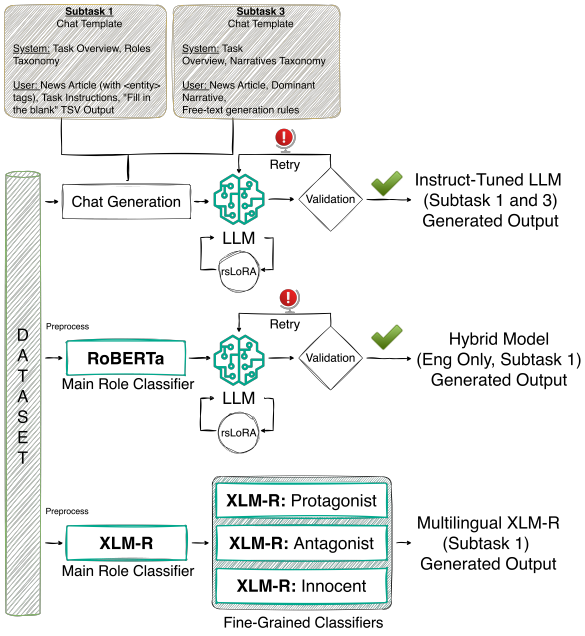

Figure 1: Schematic representation delineating the data preprocessing and model execution pipeline for Subtasks 1 and 3

both main and fine-grained roles performed best. For Subtask 3, we instruct-tuned Phi-4 to generate structured explanations, incorporating details about the article and its dominant narrative. The need for scalable solutions capable of processing millions of articles daily necessitates an approach that prioritizes efficiency and minimizes resource consumption. We aim to build a system that effectively addresses the challenges of the task while maintaining computational efficiency.

Table 1 captures the performance of our system against the baseline and other teams in the competition[1]. Our system performance is particularly competitive against the competition in Hindi (2^{nd}) and Russian (3^{rd}) for Subtask 1. We observed that few-shot prompting was suboptimal compared to LoRA-based parameter-efficient fine-tuning of the quantized Phi-4 model. RoBERTa models struggled with fine-grained role classification, necessitating an LLM-based approach. Our work is publicly available[2] for reproducibility.

2 Background

The Table 2 and Table 3 illustrate the distribution of news articles for each subtask and language, across training, validation, and test splits. Our approach is highlighted in Figure 1 and elucidated in section 3.

[1] Official Leaderboard

[2] GitHub Repo

	Subtask 1	Subtask 2	Subtask 3
Bulgarian	401 / 15 / 54	401 / 35 / 100	401 / 28 / 79
English	399 / 27 / 63	399 / 41 / 101	399 / 30 / 68
Portuguese	400 / 31 / 71	400 / 35 / 100	400 / 25 / 83
Hindi	366 / 35 / 78	366 / 35 / 99	366 / 29 / 40
Russian	215 / 28 / 57	215 / 32 / 60	215 / 28 / 56

Table 2: Distribution of news articles per subtask and language, across train / validation / test splits

	Avg Entity Count per News Article	Avg Fine-Grained Role Count per Entity
Bulgarian	2.42	2.96
English	3.40	2.20
Portuguese	4.09	2.14
Hindi	6.82	1.40
Russian	3.41	2.08
Overall	4.25	2.11

Table 3: Statistics on the Training Data - Subtask 1

2.1 Subtask 1 - Entity Framing

The subtask involves the classification of entity mentions within news articles, using a hierarchical taxonomy of roles. This presents a multi-label, multi-class text-span classification problem, as each entity mention can be assigned multiple fine-grained roles from a predefined set. The dataset exhibits complexities such as multiple entities per article, repeated occurrences of the same entity, and even multiple annotations of the same entity within a single article, potentially with varying roles. These nuances require careful consideration when developing and evaluating classification models for this task. For each annotated entity, the system must assign a single main role, followed by one or more fine-grained roles based on the main role. The complete set of main and fine-grained roles was provided as a predefined taxonomy[3] (Stefanovitch et al., 2025).

2.2 Subtask 2 - Narrative Classification

The subtask follows a hierarchical classification structure similar to Subtask 1, having coarse-grained and fine-grained narratives which was provided as a predefined taxonomy[4] (Stefanovitch et al., 2025) by the task. Finally, one of the narratives or sub-narratives is assigned as a dominant narrative at the article level. While this setup presents several interesting challenges, it falls outside the scope of the present work and is left as a

[3] https://propaganda.math.unipd.it/semeval2025task10/ENTITY-ROLE-TAXONOMY.pdf

[4] https://propaganda.math.unipd.it/semeval2025task10/NARRATIVE-TAXONOMIES.pdf

potential direction for future research.

2.3 Subtask 3 - Narrative Extraction

This subtask involves generating a concise free-text explanation that supports the dominant narrative of a news article, making it a text-to-text generation task. The explanation is based on the text fragments that justifies the claims of the dominant narrative which is essentially an output from Subtask 2.

2.4 Related Work

Various datasets have been developed to analyze narratives and sentiment in text across different domains. (Sharma et al., 2023) presents a dataset for identifying heroes, villains, and victims in memes, employing broad categories similar to the Subtask 1 dataset. In (Coan et al., 2021), a dataset is introduced for classifying Climate Change denial claims, which applies a narrative taxonomy similar to Subtask 2. Additionally, (Amanatullah et al., 2023) offers a detailed examination of narratives in the context of the Ukraine-Russia war.

Our approach primarily focuses on instruct-tuning quantized LLMs, particularly Phi-4, along with comprehensive prompt engineering. Instruct tuning, a form of supervised fine-tuning, significantly enhances LLM capabilities by aligning them with human instructions and downstream tasks (Zhang et al., 2023). Emerging techniques such as Parameter Efficient Fine Tuning (PEFT) (Fu et al., 2023), especially Low Rank Adaptation (LoRA) (Hu et al., 2022; Kalajdzievski, 2023), have made it significantly easier to instruct-tune LLMs with limited computational resources.

Recent advancements emphasize the effectiveness of instruction tuning and few-shot prompting. (Brown et al., 2020) demonstrated that few-shot prompting in LLMs (In-Context Learning paradigm) could generalize to a variety of tasks with minimal fine-tuning. Moreover, the benefits of instruction tuning for improving multilingual capabilities have been highlighted by (Chirkova and Nikoulina, 2024; Ming et al., 2024), suggesting that future LLM development should prioritize multilingual training. In addition, LLMs can tackle diverse and challenging tasks, including specialized domain understanding and complex zero-shot reasoning (Raghav et al., 2023, 2025). Furthermore, ongoing research explores critical aspects of modern models to enhance robustness (Carragher et al., 2025a,b).

Before the advent of large-scale LLMs, encoder-decoder models such as T5 and BART (Raffel et al., 2020; Lewis et al., 2019) were effective across a variety of NLP tasks (Raghav et al., 2022). Prior work in specialized domains (Mullick et al., 2022b,a, 2023) has explored fine-grained classification tasks closely related to our objectives. In our work, we explored RoBERTa for role classification in Subtask 1, as its effectiveness in text classification tasks is well-documented (Liu et al., 2019).

3 System Overview

This section presents the key algorithmic and modeling decisions behind our system. We rely exclusively on the dataset provided by the organizers and systematically examine a range of approaches, including large and small language models, In-Context Learning (ICL) paradigms (zero-shot and one-shot), and instruction-tuning methods. Due to compute constraints, we were unable to use large non-quantized models and instead relied on quantized large models and non-quantized small models. We observed that the quantized large models consistently outperformed their smaller non-quantized counterparts. Our approach leverages specialized instruct-tuned quantized LLMs, where well-crafted prompts guide the LLM, followed by lightweight fine-tuning using rsLoRA. We address challenges such as multilingual training, long-document parsing, and robust text generation through targeted strategies.

3.1 Instruct-Tuned LLM Approach

We employed instruct tuning across several LLMs, including LLaMA and Phi-4. Through experimentation, we found that Phi-4 achieved the best performance when tuned using rsLoRA. To address key challenges, we adopted the following strategies:

- **Entity Tagging:** We explicitly marked entity mentions in the text using `<entity>` tags to help the model focus on relevant spans. This helps the model distinguish between multiple occurrences of the same entity in the same news article.

- **Structured Output Format:** Instead of free-text predictions, we guided the model to produce structured Tab Separated Values (TSV) outputs of the form:

```
Entity \t Main_Role \t Fine_Grained_Role(s)
```

As we need to generate multiple fine-grained roles, we allow for generation in a comma seperated format. This prevented the model from overfitting to format structure (like JSON) rather than meaningful content. This helps optimise the loss function efficiently which can be quantitatively seen in Table 6.

- **Error Correction via Iterative Refinement:** For Subtask 1, we implement an iterative feedback mechanism in which errors from incorrect or unparsable outputs were fed back to the LLM for up to three retries, refining predictions dynamically. This process was triggered whenever the output did not conform to the expected TSV/JSON format or when the predicted fine-grained role was not consistent with the corresponding main role. Similarly for Subtask 3, to ensure clarity and control length, we imposed structural constraints and implemented an iterative retry mechanism for generated outputs exceeding 80 words.

- **Prompt Engineering:** We explored various prompt formulation strategies and adopted a System-User-Assistant chat template that yielded the best results for our task. The *System* message included the task introduction and a detailed taxonomy, while the *User* message contained the input article and generation instructions. Both zero-shot and instruction tuning approaches followed the same template. In the one-shot setting, we additionally provide an illustrative input-output example within the *User* message. A detailed illustration of the prompt can be found at subsection A.1 and subsection A.2.

3.2 Hybrid Model: RoBERTa for English Role Classification

RoBERTa works well on text classification in NLP tasks in the English language, especially when we have limited number of classes. We incorporated a RoBERTa-based classifier for main-role prediction. This model achieved a validation classification accuracy of 71.1% when distinguishing amongst the main roles ($\sim$2% improvement over LLMs). The predicted main roles were then fed into Phi-4 for fine-grained role classification.

We address RoBERTa's token limit by adopting a windowed-context approach, extracting 300 characters before and 200 after the target entity to retain critical surrounding context.

3.3 Multilingual Training with XLM-R

For multilingual support, we adopt a two-stage windowed-context based classification. First, we train a three-way classifier to predict the main role of an entity. Based on this prediction, the input is routed to one of three fine-grained multi-class classifiers, each specializing in a specific main role. While this approach beat the baseline, it achieved an exact match ratio of only 0.24 for English. Further, we observed lower performance for other languages.

4 Experimental Setup

Our approach for Subtasks 1 and 3 (Figure 1) focused on instruct-tuning quantized LLMs, enhanced by prompt engineering. We utilized the provided training split for systematic instruction tuning, the validation split for model selection and performance assessment, and the blind test split to generate predictions. To address the multilingual nature of the task, we adopt a combined training strategy, leveraging data from all languages to create a unified instruct-tuned model.

4.1 Instruct Tuning Setup

We leverage Unsloth's (Daniel Han and team, 2023) dynamic 4-bit quantized Phi-4 model[5] (8.48B parameters) as our base model for efficient LLM training. We use the Supervised Fine-Tuning Trainer (SFTT) from the Hugging Face TRL Library (von Werra et al., 2020) for instruct tuning. We also experimented with LLaMA-3.2 3B[6] and 4-bit dynamic quantized LLaMA-3.2 3B Instruct[7] from Unsloth.

We limit training to one epoch, as further training led to performance degradation. Furthermore, we integrated rsLoRA ($rank = 8$) and utilized Unsloth's gradient checkpointing. Our experiments (training and data storage) were conducted using a single T4 GPU on Google Colab and Kaggle.

4.2 Evaluation Metrics

Subtask 1 employs Exact Match Ratio, assessing the proportion of samples with perfect agreement between predicted and ground truth labels for both main and fine-grained roles. Subtask 2 utilizes samples F1 calculated per document. Subtask 3 leverages the F1 metric from BERTScore (Zhang

[5] https://huggingface.co/unsloth/phi-4-unsloth-bnb-4bit
[6] https://huggingface.co/unsloth/Llama-3.2-3B
[7] https://huggingface.co/unsloth/Llama-3.2-3B-Instruct-bnb-4bit

Model	Training Methodology	Bulgarian	English	Portuguese	Hindi	Russian
LLaMA 3.2 3B	Zero-Shot (ICL)	0.00	0.02	0.02	0.01	0.03
	One-shot (ICL)	0.02	0.02	0.03	0.02	0.05
	Instruct Tuning	0.08	0.11	0.09	0.12	0.10
LLaMA 3.2 - 3B Instruct	Zero-shot (ICL)	0.03	0.03	0.04	0.02	0.07
	One-shot (ICL)	0.04	0.05	0.05	0.04	0.07
	Instruct Tuning	0.09	0.13	0.10	0.14	0.11
Phi-4	Zero-shot (ICL)	0.25	0.22	0.40	0.31	0.32
	One-shot (ICL)	0.27	0.24	0.41	0.35	0.38
	Instruct Tuning	**0.42** (4^{th})	0.35	**0.70** (5^{th})	**0.49** (1^{st})	**0.55** (4^{th})
XLM-R (Cross-Lingual)	Fine Tuning	0.09	0.24	0.17	0.11	0.12
RoBERTa + LLM (English)	Fine Tuning	-	**0.37** (9^{th})	-	-	-

Table 4: Validation set results for Subtask 1 from the ablation study combining model variants with different training strategies. TSV is used as the output format, based on results from Table 6. Performance is measured using Exact Match Ratio.

Model	Training Methodology	Bulgarian	English	Portuguese	Hindi	Russian
LLaMA 3.2 3B	Zero-Shot (ICL)	0.18	0.19	0.21	0.26	0.24
	One-shot (ICL)	0.25	0.27	0.30	0.33	0.32
	Instruct Tuning	0.45	0.48	0.47	0.49	0.46
LLaMA 3.2 - 3B Instruct	Zero-Shot (ICL)	0.21	0.22	0.22	0.28	0.29
	One-shot (ICL)	0.29	0.31	0.32	0.37	0.38
	Instruct Tuning	0.48	0.52	0.50	0.53	0.49
Phi-4	Zero-Shot (ICL)	0.29	0.32	0.34	0.37	0.35
	One-shot (ICL)	0.37	0.39	0.44	0.49	0.59
	Instruct Tuning	**0.67** (6^{th})	**0.71** (21^{st})	**0.70** (6^{th})	**0.70** (4^{th})	**0.68** (3^{rd})

Table 5: Validation set results for Subtask 3 from the ablation study combining model variants with different training strategies. Performance measured on BERTScore F1.

et al., 2019) to measure the similarity between generated and gold explanations.

5 Results

Our results (Table 1) demonstrate the effectiveness of prompt engineering and instruct-tuned LLMs for multilingual entity framing and narrative extraction.

We conduct two sets of ablations for Subtask 1: model variant combined with output format (TSV vs. JSON) as in Table 6, and model variant combined with training strategy (ICL zero-shot, ICL one-shot, and instruct tuning) as in Table 4. For Subtask 3, we similarly evaluated model variants with different training strategies to systematically improve performance as in Table 5. These studies helped isolate the effects of output format, training methodology and model variant effectiveness. Our key findings include:

- **LLaMA 3.2** models perform poorly for both entity classification and free-text generation. Overall, instruct tuning with **Phi-4** yields the best results.
- The instruct model variants (such as LLaMA 3.2 - 3B Instruct), consistently outperform their base counterparts (such as LLaMA 3.2 3B) even after instruct-tuning.
- We tried out zero-shot and one-shot ICL prompting methods but both these methods were worse than instruct tuning.
- Iterative **retry methodologies** reduced unparsable outputs and hallucinations, while improving response validity.
- For entity framing in English, our RoBERTa+LLM model achieved an **37% exact match**. However, multilingual generalization posed a serious challenge.
- **TSV-based outputs** improved accuracy over JSON outputs, as loss values were no longer artificially lowered by structural correctness.

6 Conclusion

Our system demonstrates the effectiveness of prompt engineering and instruct-tuned quantized LLMs for multilingual entity framing and narrative extraction. We highlight the superior performance of quantized LLMs and the benefits of incorporating a hybrid model with RoBERTa for main role

prediction in English. Additionally, we emphasize the positive impact of iterative refinement and structured output formats on overall accuracy.

Future research will focus on incorporating Subtask 2 as pre-training for Subtask 3, to refine narrative extraction quality. Investigation into the impact of model size and alternative efficient fine-tuning techniques could also yield valuable insights. Finally, enhancing multilingual generalization for RoBERTa and XLM-R may yield better results.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Samy Amanatullah, Serena Balani, Angela Fraioli, Stephanie M McVicker, and Mike Gordon. 2023. Tell us how you really feel: Analyzing pro-kremlin propaganda devices & narratives to identify sentiment implications. *The Propwatch Project. Illiberalism Studies Program Working Paper. https://www. illiberalism. org/tell-us-how-you-really-feel-analyzing-pro-kremlinpropaganda-devices-narratives-to-identify-sentiment-implications*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Peter Carragher, Abhinand Jha, R Raghav, and Kathleen M Carley. 2025a. Quantifying memorization and retriever performance in retrieval-augmented vision-language models. *arXiv preprint arXiv:2502.13836*.

Peter Carragher, Nikitha Rao, Abhinand Jha, R Raghav, and Kathleen M Carley. 2025b. Koala: Knowledge conflict augmentations for robustness in vision language models. *arXiv preprint arXiv:2502.14908*.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. *arXiv preprint arXiv:2402.14778*.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12799–12807.

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.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Jones, Jingfei Liu, David Romero, Lucas Gracia, Denis St-Amand, Lori Shaw, et al. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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. *arXiv preprint arXiv:1907.11692*.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, et al. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *arXiv preprint arXiv:2412.04003*.

Ankan Mullick, Ishani Mondal, Sourjyadip Ray, Raghav R, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in Indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1870–1881, Dubrovnik, Croatia. Association for Computational Linguistics.

Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, Raghav R, and Roshni Kar. 2022a. An evaluation framework for legal document summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753, Marseille, France. European Language Resources Association.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022b. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551.

R Raghav, Jason Rauchwerk, Parth Rajwade, Tanay Gummadi, Eric Nyberg, and Teruko Mitamura. 2023. Biomedical question answering with transformer ensembles. In *CLEF (Working Notes)*.

R Raghav, Adarsh Vemali, and Rajdeep Mukherjee. 2022. Etms@ iitkgp at semeval-2022 task 10: Structured sentiment analysis using a generative approach. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1373–1381.

R Raghav, Adarsh Prakash Vemali, Darpan Aswal, Rahul Ramesh, and Ayush Bhupal. 2025. Scotty-poseidon at semeval-2025 task 8: Llm-driven code generation for zero-shot question answering on tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? *arXiv preprint arXiv:2301.11219*.

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).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

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

A Appendix

A.1 Prompt Template - Subtask 1

In Subtask 1, our strategy follows the system-user-assistant chat paradigm, a structured format to ensure accurate entity classification. In the 'System' prompt, we provide a taxonomy defining the main and fine-grained roles. This is essential as the task definitions of these entities are slightly different than what they may mean in general settings. In the 'User' prompt, goes the article text, modified with <entity> tags to identify entity instances that need classification. We add clear instructions to the model for generating classifications at both role levels, along with a reiteration to emphasize on the target main and fine-grained roles. We provide a fill-in-the-blank structure for every instance of entities using a tab-separated format to improve model consistency and generation accuracy.

System

```
You are an expert assistant trained for fine-grained entity classification. Your purpose is to
assign accurate roles to tagged entities in an article based on predefined categories.
For each tagged entity, there can be one main role from this list - ['Protagonist', 'Antagonist',
'Innocent']

Here are the fine-grained roles for Protagonist
<fine-grained roles and its descriptions for Protagonist>

Here are the fine-grained roles for Antagonist
<fine-grained roles and its descriptions for Antagonist>

Here are the fine-grained roles for Innocent
<fine-grained roles and its descriptions for Innocent>
```

User

```
User Prompt
### Task Instructions

#### Article:
<the news article, with <entity> tags>

For every tagged entity phrase, you must:
1. Assign *one main role* from the list: ['Protagonist', 'Antagonist', 'Innocent'].
2. Assign *one or more fine-grained roles* based on the predefined taxonomy for each main role:
   - *Protagonist*: Guardian, Martyr, Peacemaker, Rebel, Underdog, Virtuous.
   - *Antagonist*: Instigator, Conspirator, Tyrant, Foreign Adversary, Traitor, Spy, Saboteur,
   Corrupt, Incompetent, Terrorist, Deceiver, Bigot.
   - *Innocent*: Forgotten, Exploited, Victim, Scapegoat.
3. The output should contain tab 3 sets separated values - entity, main role and fine grained roles
4. Fine grained roles should be separated by comma if there are multiple fine grained roles for the
entity

Fill in the missing information in the following structure:
<A TSV/JSON structure containing the actual entity name, along with placeholders such as
`main_role`, and `fine_grained_role1, fine_grained_role2` to represent the main and fine-grained
roles, respectively>
The output should be the filled up.
```

Assistant

The 'Assistant' role is the output of the model, which was provided during training. Below is an example of the assistant output for one of the training examples (the news article `EN_UA_300009.txt`) in English, presented in TSV format.

```
Fail Alsynov    Protagonist   Rebel, Martyr
Bashkir people  Innocent      Victim
Bashkort        Protagonist   Rebel, Guardian
```

Here is the same assistant output in JSON format.

```
{
    'Fail Alsynov':{
        'main_role': 'Protagonist',
        'fine_grained_role': ['Rebel', 'Martyr']
    },
    'Bashkir people':{
        'main_role': 'Innocent',
        'fine_grained_role': ['Victim']
    },
    'Bashkort':{
        'main_role': 'Protagonist',
        'fine_grained_role': ['Rebel', 'Guardian']
    }
}
```

A.2 Prompt Template - Subtask 3

In Subtask 3, the tiered taxonomy is included in the System prompt, while the User prompt contains the topic label (e.g., 'Climate Change', 'Ukraine–Russia War', or 'Other'), the coarse and fine-grained narratives, and the article text. We also provide clear instructions to generate free-text explanations for generating justifications of the narratives.

System

```
You are an expert assistant trained for generating a free text explanation given the narrative and
the subnarrative of the article.
Given a news article and a dominant narrative of the text of this article, you should generate a
free-text explanation supporting the choice of this dominant narrative. The to-be-generated
explanation should be grounded in the text
fragments that provide evidence of the claims of the dominant narrative.

These are the definitions for narratives and subnarratives which you will encounter for the text:
<coarse narratives, their fine-grained narratives and their descriptions listed respectively
```

User

```
### Task Instructions

#### Article:
This is the article from which you need to extract the explanation of why the narrative and
subnarrative make sense.
<the raw news article>

Here are the categories, dominant narratives and subnarratives from the text that need an
explanation of why they are so.
Category: <`Climate Change`, `Ukraine Russia War` or `Other`>
Narrative: <coarse-grained narrative>
Subnarrative: <fine-grained narrative, if present>

Generate a one-paragraph explanation in the same language as the provided article, strictly limited
to **three sentences or 60 words**, whichever comes first. This requirement is **absolute and
non-negotiable**.
```

Assistant

The 'Assistant' role is the output of the model, which was provided during training. Below is an example of the assistant prompt for one of the training examples (the news article `EN_CC_100013.txt`) in English.

```
The text accuses climate activist Bill Gates for his alleged hypocritical behavior as he flies in
private jets that pollute the environment while advocating for the climate cause.
```

A.3 Ablation Results

Model	Output Format	Bulgarian	English	Portuguese	Hindi	Russian
LLaMA 3.2 3B	JSON	0.00	0.02	0.01	0.01	0.02
	TSV	0.00	0.02	0.02	0.01	0.03
LLaMA 3.2 - 3B Instruct	JSON	0.02	0.02	0.03	0.02	0.05
	TSV	0.03	0.03	0.04	0.02	0.07
Phi-4 Instruct	JSON	0.18	0.20	0.33	0.23	0.22
	TSV	0.25	0.22	0.40	0.31	0.32

Table 6: Ablation study on the validation set for Subtask 1 to compare output formats (JSON vs. TSV). All models were evaluated using Exact Match Ratio. While LLaMA models showed negligible differences across formats, Phi-4 demonstrated a significant improvement with TSV, leading us to adopt TSV as the preferred output format.