DEMON at SemEval-2025 Task 10: Fine-tuning LLaMA-3 for Multilingual Entity Framing

Matteo Fenu and **Manuela Sanguinetti** and **Maurizio Atzori**
Department of Mathematics and Computer Science, University of Cagliari
Via Ospedale 72, Cagliari (Italy)
m.fenu28@studenti.unica.it,{manuela.sanguinetti,atzori}@unica.it

Abstract

This study introduces a methodology centred on Llama 3 fine-tuning for the classification of entities mentioned within news articles, based on a predefined role taxonomy. The research is conducted as part of SemEval-2025 Task 10, which focuses on the automatic identification of narratives, their classification, and the determination of the roles of the relevant entities involved. The developed system was specifically used within Subtask 1 on Entity Framing. The approach used is based on parameter-efficient fine-tuning, in order to minimize the computational costs while maintaining reasonably good model performance across all datasets and languages involved. The model achieved promising results on both the development and test sets. Specifically, during the final evaluation phase, it attained an average accuracy of 0.84 on the main role and an average Exact Match Ratio of 0.41 in the prediction of fine-grained roles across all the five languages involved, i.e. Bulgarian, English, Hindi, Portuguese and Russian. The best performance was observed for English (3rd place out of 32 participants), on a par with Hindi and Russian. The paper provides an overview of the system adopted for the task and discusses the results obtained.

1 Introduction

The way entities are presented within a text plays a crucial role in shaping the narrative and influencing public opinion. Entity framing refers to the process by which a text assigns specific roles to the actors involved in an event, based on a predefined set of roles. This phenomenon is far from neutral, as defining a subject as a "victim" rather than a "perpetrator," for example, can significantly influence how an event is perceived by the reader. The study of the dynamics of mis/disinformation and information manipulation has received growing attention, both within the NLP community and beyond (Wardle, 2018), and understanding how language structures reality and public debate has become a key task. Several efforts have been made in this direction, both in developing computational–though theoretically-grounded–frameworks (Minnema et al., 2022; Wang et al., 2024) and in creating linguistic resources and label taxonomies aimed at thoroughly analyzing the phenomenon (Ziems and Yang, 2021; Mahmoud et al., 2025). These resources, in turn, contribute to advancing the state of the art in automatic approaches.

In this paper, we present a computational approach to this challenging issue in the context of our participation in Subtask 1 of SemEval-2025 Task 10 on Multilingual Characterization and Extraction of Narratives from Online News (Piskorski et al., 2025). Subtask 1 on Entity Framing precisely focuses on assigning one or more roles to entities mentioned within an article, based on a predefined taxonomy. This task is proposed as a span classification problem and presents both multi-class and multi-label challenges: multiple roles can be assigned to a single entity, and the number of possible roles is extensive. Our approach relies on parameter-efficient fine-tuning using QLoRA and Llama 3 (8B parameters) as the reference model. During training, we experimented with different prompting strategies in order to assess the impact on results of two key factors:

- The number of fine-grained roles predicted by the model (this aspect in particular is related to the challenges posed by multi-label classification).
- The influence of surrounding context.

The paper provides an overview of the task addressed and the approach followed in developing our system, finally discussing the results obtained on the datasets released for the competition in the different setups.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1456–1464

2 Background

Given a news article and a list of entity mentions (including their span offsets), the task of entity framing consists in assigning one or more roles to each entity based on a predefined taxonomy. The taxonomy provided for this task covers three main roles: *protagonist*, *antagonist* and *innocent*. Protagonists represent entities with a positive role in society, who actively strive for the common good. In contrast, antagonists oppose the good deeds of the protagonists by performing cruel acts against people or things. In between these two opposing factions are the innocents. They represent the outcasts and people who are victims of injustice. A detailed list of fine-grained roles is associated with each one of these main roles.[1]

The dataset made available by the organizers consists of articles dealing with two main topics, Ukraine-Russia war and climate change, and it includes five languages: English, Bulgarian, Hindi, Russian and Portuguese. Each article in the dataset includes a title and its content, while the annotations specify entity classifications. For each entity, the dataset provides its position within the article, the assigned main role along with the associated fine-grained roles.

Concerning the task evaluation criteria, in addition to metrics assessing main role accuracy and measures such as micro-precision, micro-recall, and micro-F score, the primary evaluation metric used to determine the ranking is the Exact Match Ratio (EMR), which is computed as follows:

$$\text{EMR} = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$

where n is the total number of samples, I is the indicator function, which is equal to 1 if the predicted value matches the true value and 0 otherwise. This is a highly stringent metric, as it prioritizes fully correct results while penalizing partially correct ones.

Next section describes the main characteristics of the system, which takes into account the task's peculiarities just outlined.

[1]Due to space constraints, we do not include in this overview the whole set of fine-grained roles, that can instead be consulted at the following link: https://propaganda.math.unipd.it/semeval2025task10/ENTITY-ROLE-TAXONOMY.pdf.

3 Dataset Overview

The dataset used for the model training phase consists of approximately 5500 annotations, in which in 47% of the cases the annotated entities are classified as antagonists, in 32% of the cases as protagonists, and in 21% of the cases as innocents. An imbalance towards the antagonist class is therefore immediately apparent, and is also visible in the evaluation dataset, which justifies more annotations on this class in order to optimize the classifier. With regard to subclass labels, the distribution is also uneven. In fact, some labels have a significantly low number of annotations. In the case of "Spy" or "Martyr", for instance, the number of annotations is less than 1%. More details on the dataset development and composition are provided in the main task report (Piskorski et al., 2025).

4 System Overview

The pipeline followed for the system development included three main steps of data pre-processing, prompt definition and actual fine-tuning. As further detailed below, the former two steps aimed at properly addressing the challenges posed by multi-label classification and the impact of different context windows, while the fine-tuning process was set so as to reduce the computational cost deriving from the use of large language models.

Data pre-processing This phase begins with the segmentation of the input articles into individual sentences using a pre-trained model tailored to each source language. Afterward, the tokenized text is processed by a module that identifies the relevant portion of the article containing the entity to be classified, alongside its annotation (entity name and character offset). A variable parameter, the context size, is specified, to determine the number of sentences included in the final passage to pass to the prompt. Furthermore, non-English data was translated into English.

At this stage, a comparison was made against a number of Python libraries for the automatic translation of sentences, from which it emerged that GoogleTrans [2] offered a good trade-off between response time and translation quality.

The last operation performed in this phase is the augmentation of the training dataset, in order to extend the distribution of underrepresented sub-roles. As mentioned in Section 3, the training dataset is

[2]https://pypi.org/project/googletrans/

characterized by a strong imbalance in the distribution of sub-class labels. New examples were generated from existing ones, through the use of the Llama 3-8b model and the definition of a simple prompt. The instructions given to the model include the replacement of some terms with synonyms, leaving the name of the entity and the general meaning of the sentence unchanged.During this phase, a series of tests are carried out to ensure good quality in the examples generated via LLM. In particular, from a set of elements generated by Llama via the starting prompt, a manual analysis of the results is performed to identify potential frequent errors from which additional prompts can be generated. One of the most common errors concerns the fact that the model tries to achieve the result by substituting names of persons or things. To try to refine the result, synonyms are specified in the prompt. During the actual data augmentation process, a check is performed on the output produced by the model to ensure the presence of the entity to be classified within the sentence. If this check fails, the generated sentence is discarded in order to avoid the addition of noise within the final dataset.

Prompt definition The model training process relies on a fine-tuning technique that uses prompts to represent annotations. A prompt is structured defining the classification problem and listing the possible sub-classes that can be assigned to an entity. The prompt then includes the entity's name, the relevant article portion, and the correct label for the entity. During evaluation, the model is provided with the same prompt, where it must predict the appropriate sub-classes. A crucial aspect of the approach involves handling annotations associated with multiple sub-roles. To address this, we devised three possible prompting strategies:

- *S1 - Single prompt with all fine-grained roles*: The prompt precisely includes a single example with all the associated roles and sub-roles. The underlying assumption is that training the model with a single prompt per full annotation should be beneficial, as it exposes the model to the full set of expected sub-classes for each entity. This setup is intended to help the model learn the relationships between subclasses, hence maximizing the EMR.

- *S2 - Different prompts for each fine-grained role*: The annotation is split into multiple examples, each featuring only one sub-role. This approach aims at maximizing the model's ability to recognize individual sub-roles independently and at reducing the complexity of each training instance.

- *S3 - Mixed approach*: It combines elements of the previous strategies, with the aim of balancing their advantages.

While experiments were carried out with all three strategies, as also discussed in Section 6.1, S2 was eventually selected as primary prompting strategy.

Fine-tuning The model is fine-tuned using the QLoRA technique (Dettmers et al., 2024), which combines quantization with LoRA. Specifically, QLoRA applies quantization to reduce memory requirements while employing LoRA to adapt the model's parameters via low-rank matrices. The adaptation is controlled using key configuration parameters, including $lora\_alpha$, $lora\_dropout$, and r, which regulate the scaling of the low-rank matrices, dropout probability, and the rank of the adaptation matrices, respectively.

5 Experiment Setup

All the experiments carried out for this task were performed only using the data made available by the organizers. As mentioned in Section 4, during the training phase, we augmented the data, thus passing from an overall amount of around 5500 to 6750 annotated entities across the five languages.

For the data pre-processing step, we used Stanza (Qi et al., 2020) for the sentence splitting (as this library supports all the five languages included in the dataset) and the GoogleTrans library to translate the data in English. As regards the context window, we observed that increasing the number of sentences to include in the input did not necessarily lead to better predictions. As a matter of fact, this often introduced conflicting annotations for the same entity. After several testings, we finally opted for a span of three sentences, as this allowed to provide a reasonable amount of information to get accurate predictions.

The model used is Meta-Llama-3-8B-Instruct[3]. For its fine-tuning, the model was loaded with four-bit quantization, and the LoRA parameters

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

were configured as follows: lora_aplha = 16, lora_dropout=0 and r=6. AdamW was used as optimizer. The model was trained over five epochs due to the small dataset size. With particular reference to the QLoRA, the training settings used are: gradient_accumulation_steps=8, learning_rate=2e-4, max_grad_norm=0.3, warmup_ratio=0.03, per_device_train_batch_size=1, loss_function = cross-entropy.

All experiments were performed in a Google Colab environment, with a A100 GPU (40 GB VRAM).

6 Results

This section aims to provide an overview not only of the final results obtained during the evaluation phase, but also of the preliminary results obtained in the development phase while experimenting with the different prompting strategies.

6.1 Results on the Development Set

As described in Section 4, different approaches were followed to prepare the data for the fine-tuning process, with the aim of assessing which one would better address the challenges deriving from both the extensive set of available fine-grained roles and the possibility of assigning multiple classes to each relevant entity within the articles. For the classification of the main role, it is assigned automatically based on the predicted subclass membership.

As regards S1 strategy, where the prompt includes a single example with all the associated roles and sub-roles, despite some encouraging results on the Portuguese set, the approach proved unsuccessful overall, as also shown in Table 1. While, in fact, the model was generally able to properly identify the main role (as shown by the average accuracy), correctly predicting all sub-roles in one pass consistently proved more challenging in all languages, especially in Bulgarian and English.

Lang.	EMR	Micro P	Micro R	Micro F1	Accuracy Main Role
EN	0.3846	0.4494	0.40	0.4233	0.9231
PT	**0.6034**	0.6552	0.6129	0.6333	0.9052
BG	0.3871	0.4839	0.4412	0.4615	0.8065
HI	0.4536	0.5236	0.4675	0.4940	0.8036
RU	0.50	0.5116	0.4944	0.5029	0.907
avg.	0.4657	0.5247	0.4832	0.5030	0.869

Table 1: Results obtained on the development set with the S1 prompting strategy.

In the alternative approach (i.e., S2), each annotation involving several sub-classes was divided into a number of prompts corresponding to the number of sub-classes assigned. This strategy resulted in significant improvements in terms of main role accuracy and EMR for English and Portuguese; for the remaining languages conflicting behaviors were observed: while the accuracy for the main role decreased, the prediction of sub-roles actually benefited from this kind of approach and resulted in a consistent increase of micro-P/R/F_1 and EMR, as reported in Table 2.

Lang.	EMR	Micro P	Micro R	Micro F1	Accuracy Main Role
EN	0.5385	0.5652	0.52	0.5417	0.956
PT	**0.7069**	0.7672	0.7177	0.7417	0.9655
BG	0.4194	0.5161	0.4706	0.4923	0.7419
HI	0.475	0.5487	0.4935	0.5197	0.7929
RU	0.5349	0.5465	0.5281	0.5281	0.8953
avg.	**0.5349**	**0.5887**	**0.5459**	**0.5647**	**0.8703**

Table 2: Results obtained on the development set with the S2 prompting strategy.

The third strategy tested attempts to combine the two previous approaches, showing the model both multi-class prompts and individual predictions. The full results are reported in Table 3. Similarly to the previous results, the highest scores are obtained on the Portuguese data; however, when comparing such values to the ones obtained with S1 and S2, we observe that this mixed approach did not contribute to the model's improvement, neither in terms of main role prediction (with an accuracy value comparable to S1) nor of sub-roles. While in fact the remaining scores are higher than the ones in S1, they do not outperform S2. We thus remark that, in these settings, predicting multiple sub-roles at once can be more penalizing, especially in terms of EMR, compared to predicting a single label. This motivated our choice to finally use the model fine-tuned with the S2 approach to submit our predictions for the final evaluation phase.

Lingua	EMR	Micro P	Micro R	Micro F1	Accuracy Main Role
EN	0.4505	0.5056	0.4500	0.4762	0.9121
PT	**0.6638**	0.7241	0.6774	0.7000	0.9397
BG	0.3871	0.4839	0.4412	0.4615	0.7742
HI	0.4500	0.5233	0.4740	0.4974	0.7821
RU	0.5000	0.5116	0.4944	0.5029	0.9186
avg.	0.5073	0.5497	0.5074	0.5276	0.8653

Table 3: Results obtained on the development set with the S3 prompting strategy.

As additional experiment in this phase, we further tested the S2 strategy with the aim of assessing the impact of the context dimension on the per-

formance of the model. Table 4 shows the results obtained from the fine-tuned model on a dataset of annotations in which only one context sentence is taken for classification, unlike the previous experiments in which the dataset was developed using a context size of three sentences (as also mentioned in Section 5).

Lingua	EMR	Micro P	Micro R	Micro F1	Accuracy Main Role
EN	0.4505	0.5056	0.4500	0.4762	0.9121
PT	**0.6638**	0.7241	0.6774	0.7000	0.9397
BG	0.3871	0.4839	0.4412	0.4615	0.7742
HI	0.4500	0.5233	0.4740	0.4974	0.7821
RU	0.5000	0.5116	0.4944	0.5029	0.9186
avg.	0.5073	0.5497	0.5074	0.5276	0.8653

Table 4: Results obtained on the development test using a single context sentence.

As expected, reducing the context window resulted in lower performance overall, thus confirming that a larger context size generally allows the model to make more accurate predictions.

6.2 Results on the Test Set

The test set provided by the organizers for Subtask 1 consists of 235 annotated entities for English, 124 for Bulgarian, 316 for Hindi, 297 for Portuguese and 214 for Russian. Table 5 reports the results obtained with S2 prompting strategy for the fine-tuning of Llama 3 on the official test set of the task. These results are also available on the official page of the task.[4]

Lang.	EMR	Micro P	Micro R	Micro F1	Accuracy Main Role
EN	0.3745	0.4487	0.3962	0.4208	0.9191
PT	0.3670	0.4324	0.3963	0.4136	0.8081
BG	0.4597	0.4797	0.4609	0.4701	0.8871
HI	0.4019	0.5253	0.4346	0.4756	0.7563
RU	**0.4673**	0.5142	0.4802	0.4966	0.8131
avg.	0.4140	0.4800	0.4336	0.4550	0.8367

Table 5: System's results on the test set for Subtask 1.

The model's highest EMR is achieved on Russian, followed by Bulgarian and Hindi; quite surprisingly, the lowest EMR score is for Portuguese, which instead was the language with the best results on the development set with all the fine-tuning approaches explored in this work. Even the values of micro-P/R/F_1 generally align with EMR, showing that the model performed best in distinguishing fine-grained roles particularly in Russian and Bulgarian, with the other languges lagging behind. As regards the main role classification, the system achieves reasonably good results across all languages, with an overall average accuracy of 0.84. Contrarily to fine-grained roles, the highest performance is obtained with the English data, while the lowest is with Hindi. This suggests that, despite the observed challenges with sub-role identification, the model remains reliable when tasked with main role categorization.

Table 5 highlights a substantial difference in performance between the development and the test set, particularly for the Portuguese language. A plausible factor contributing to this discrepancy might lie in the distribution of sub-class labels within the datasets. In the development set, the most frequent sub-class is Victim, accounting for 48% of the instances. The performance obtained by the model on this dataset suggests a good understanding of this label by Llama. Notably, Victim is also the most represented sub-class in the training data, with a frequency of 17.45% across all languages. This strong imbalance in the Portuguese development set, which favors a sub-class the model appears to properly identify, may have contributed to the high EMR observed. However, since gold labels are not available for the test set, no definitive conclusions can be drawn regarding class distribution in that partition.

7 Error Analysis

To complete our description, we carried out an exploratory error analysis of the model configuration that obtained the best results during the development phase and was finally employed for the evaluation phase; specifically, the configuration is the one featuring 3 context sentences in the prompt and using S2 as prompting strategy. The analysis was performed on the development data itself, due to the absence of the gold labels for the test set. As a case study, we opted for the Hindi section of the task dataset, since it provides a larger number of annotated entities (280) compared to the other languages, thus allowing for a more reliable basis for observing the model's behavior. The full confusion matrix is shown in Appendix B.

The sub-roles for which the model was completely unable to make correct predictions are 'Guardian', 'Traitor', 'Rebel', 'Scapegoat', 'Bigot' and 'Spy'. Conversely, the three sub-classes with the highest number of correct predictions are 'Sabo-

[4] https://propaganda.math.unipd.it/semeval2025task10/leaderboard.html

teur', 'Exploited' and 'Peacemaker'. Among these three, 'Peacemaker' stands out with an accuracy of around 0.7. Notably, 'Peacemaker' appears only in 6% of the training dataset, suggesting that label frequency alone does not determine classification success. This is further supported by the case of 'Foreign Adversary', which, despite having a frequency of over 10% in the training set, was classified with an accuracy of only 0.38.

A significant source of error involves the 'Virtuous' sub-role. As the matrix shows, the model frequently confuses this sub-class with others. In particular, the number of false positives for this sub-role is 16. A recurrent misclassification is between 'Exploited' and 'Virtuous', with the model incorrectly predicting the latter for the former 10 times. This type of error is particularly critical as it affects not only the sub-role but also the broader classification of the main role (since 'Exploited' is a specification of the 'Innocent' role, while 'Virtuous' falls under the 'Protagonist' category). The 'Exploited' sub-role proved to be quite problematic indeed, as despite achieving 27 correct predictions, it also exhibited a high number of false positives (19) and false negatives (28), indicating substantial confusion in distinguishing it from similar categories.

Certainly the main factor related to the low EMR score concerns the nature of the metric itself, as it does not distinguish partially right answers from wrong answers. This makes the correct prediction of multi-subclass annotations particularly complex. In addition, the choice of the S2 prompting strategy for training the model results in the inability of the model to predict annotations from two or more sub-classes. In spite of this, it is preferred over the other two strategies because of the greater accuracy in annotations from only one sub-class. From a grammatical point of view, the model makes several errors that can be traced back to certain writing techniques commonly used within articles; these are not correctly understood by the model. For example, passive forms often lead the model to treat entities as active subjects, despite the fact that they are the patients, i.e. the entities undergoing the action. In Example 1 below, the model mistakes Ukraine for an active subject by misclassifying it as Conspirator.

(1) *'This is a perfect example of how censorship leads to destruction. Zelensky wants Ukraine to be destroyed. There is nothing to hide'*

In some cases, one can see how the irony used by the writers leads the model to a semantic misreading. In Example 2, the model literally interprets the sentence by classifying the entity as deceiver, whereas the correct classification is Tyrant.

(2) *'Klaus Schwab wants to ban people from washing their trousers more than once a month Klaus Schwab's World Economic Forum (WEF) has issued guidelines on how often the public should be allowed to wash their clothes, including underwear and gym clothes'*

In addition, the size of the context significantly influences the correct prediction of the entities. In fact, although three article sentences are extracted as context for each classification, in some cases the sentences are short and of little meaning, thus making the classification more challenging.

8 Conclusions

The paper described an approach based on Llama 3 fine-tuning to tackle Subtask 1 on Entity Framing. The results we obtained especially within the final evaluation phase indicate that while the model generally classifies entities' main roles quite effectively, it struggles more with exact sub-role matching, as seen in the moderate scores obtained in terms of EMR, which was also the primary metric used for this task and determining its ranking. Another general remark concerns the fact that performance greatly varied by language and especially between the development and test sets, suggesting that factors such as dataset composition, but also translation effects could have had an impact on results. Future improvements could focus on enhancing sub-role differentiation, possibly through better prompting strategies or alternative fine-tuning approaches.

Code availability

The code used for the experiments described in this paper is available here: https://github.com/demon-prin/multilingual-entity-framing-of-online-news/

Acknowledgements

The work has been partially supported by the project DEMON "Detect and Evaluate Manipulation of ONline information" funded by MIUR

under the PRIN 2022 grant 2022BAXSPY (CUP F53D23004270006, NextGenerationEU), and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU (NextGenerationEU).

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLORA: Efficient Finetuning of Quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. SocioFillmore: A tool for discovering perspectives. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 240–250, Dublin, Ireland. Association for Computational Linguistics.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Guan Wang, Rebecca Frederick, Jinglong Duan, William Wong, Verica Rupar, Weihua Li, and Quan Bai. 2024. Detecting misinformation through framing theory: the frame element-based model.

Claire Wardle. 2018. The need for smarter definitions and practical, timely empirical research on information disorder. *Digital Journalism*, 6(8):951–963.

Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Prompt Examples

A.1 Example of S1 - Single prompt with all fine-grained roles

Classify the entity using one or more labels choosing from these: 'Instigator', 'Conspirator', 'Tyrant', 'Foreign Adversary',
'Traitor', 'Spy', 'Saboteur', 'Corrupt', 'Incompetent', 'Terrorist', 'Deceiver','Bigot', 'Guardian', 'Martyr','Peacemaker', 'Rebel', 'Underdog', 'Virtuous', 'Forgotten', 'Exploited', 'Victim', 'Scapegoat'.
Entity: Vladimir Putin
Context: Putin called the withdrawal of the Ukrainian Armed Forces from the Donbass and Novorossia a condition for peace 'In order to complete the special military operation (SVO) in Ukraine, Kiev must begin to implement Russia's peace initiatives, which were outlined during the Russian leader's meeting with the leadership of the Ministry of Foreign Affairs'. This statement was made by Russian President Vladimir Putin on 5 July at a press conference after negotiations with Hungarian Prime Minister Viktor Orban.
Label: Guardian, Peacemaker

A.2 Example of S2 - Different prompts for each fine-grained role

A.2.1 1st prompt

Classify the entity using one or more labels choosing from these: 'Instigator', 'Conspirator', 'Tyrant', 'Foreign Adversary',
'Traitor', 'Spy', 'Saboteur', 'Corrupt', 'Incompetent', 'Terrorist', 'Deceiver','Bigot', 'Guardian', 'Martyr','Peacemaker', 'Rebel', 'Underdog', 'Virtuous', 'Forgotten', 'Exploited', 'Victim', 'Scapegoat'.
Entity: Vladimir Putin
Context: Putin called the withdrawal of the Ukrainian Armed Forces from the Donbass and Novorossia a condition for peace 'In order to complete the special military operation (SVO) in Ukraine, Kiev must begin to implement Russia's peace initiatives, which were outlined during the Russian leader's meeting with the leadership of the Ministry of Foreign Affairs'. This statement was made by Russian President Vladimir Putin on 5 July at a press conference after negotiations with Hungarian Prime Minister Viktor Orban.
Label: Guardian

A.2.2 2nd prompt

Classify the entity using one or more labels choosing from these: 'Instigator', 'Conspirator', 'Tyrant', 'Foreign Adversary',
'Traitor', 'Spy', 'Saboteur', 'Corrupt', 'Incompetent', 'Terrorist', 'Deceiver','Bigot', 'Guardian', 'Martyr','Peacemaker', 'Rebel', 'Underdog', 'Virtuous', 'Forgotten', 'Exploited', 'Victim', 'Scapegoat'.
Entity: Vladimir Putin
Context: Putin called the withdrawal of the Ukrainian Armed Forces from the Donbass and Novorossia a condition for peace 'In order to complete the special military operation (SVO) in Ukraine, Kiev must begin to implement Russia's peace initiatives, which were outlined during the Russian leader's meeting with the leadership of the Ministry of Foreign Affairs'. This statement was made by Russian President Vladimir Putin on 5 July at a press conference after negotiations with Hungarian Prime Minister Viktor Orban.
Label: Peacemaker

A.3 Example of S3 - Mixed approach

A.3.1 1st prompt

Classify the entity using one or more labels choosing from these: 'Instigator', 'Conspirator', 'Tyrant', 'Foreign Adversary',
'Traitor', 'Spy', 'Saboteur', 'Corrupt', 'Incompetent', 'Terrorist', 'Deceiver','Bigot', 'Guardian', 'Martyr','Peacemaker', 'Rebel', 'Underdog', 'Virtuous', 'Forgotten', 'Exploited', 'Victim', 'Scapegoat'.
Entity: Vladimir Putin
Context: Putin called the withdrawal of the Ukrainian Armed Forces from the Donbass and Novorossia a condition for peace 'In order to complete the special military operation (SVO) in Ukraine, Kiev must begin to implement Russia's peace initiatives, which were outlined during the Russian leader's meeting with the leadership of the Ministry of Foreign Affairs'. This statement was made by Russian President Vladimir Putin on 5 July at a press conference after negotiations with Hungarian Prime Minister Viktor Orban.
Label: Guardian

A.3.2 2nd prompt

Classify the entity using one or more labels choosing from these: 'Instigator', 'Conspirator', 'Tyrant', 'Foreign Adversary',
'Traitor', 'Spy', 'Saboteur', 'Corrupt', 'Incompetent', 'Terrorist', 'Deceiver','Bigot', 'Guardian', 'Martyr','Peacemaker', 'Rebel', 'Underdog', 'Virtuous', 'Forgotten', 'Exploited', 'Victim', 'Scapegoat'.
Entity: Vladimir Putin
Context: Putin called the withdrawal of the Ukrainian Armed Forces from the Donbass and Novorossia a condition for peace 'In order to complete the special military operation (SVO) in Ukraine, Kiev must begin to implement Russia's peace initiatives, which were outlined during the Russian leader's meeting with the leadership of the Ministry of Foreign Affairs'. This statement was made by Russian President Vladimir Putin on 5 July at a press conference after negotiations with Hungarian Prime Minister Viktor Orban.
Label: Peacemaker

A.3.3 3rd prompt

Classify the entity using one or more labels choosing from these: 'Instigator', 'Conspirator', 'Tyrant', 'Foreign Adversary',
'Traitor', 'Spy', 'Saboteur', 'Corrupt', 'Incompetent', 'Terrorist', 'Deceiver','Bigot', 'Guardian', 'Martyr','Peacemaker', 'Rebel', 'Underdog', 'Virtuous', 'Forgotten', 'Exploited', 'Victim', 'Scapegoat'.
Entity: Vladimir Putin
Context: Putin called the withdrawal of the Ukrainian Armed Forces from the Donbass and Novorossia a condition for peace 'In order to complete the special military operation (SVO) in Ukraine, Kiev must begin to implement Russia's peace initiatives, which were outlined during the Russian leader's meeting with the leadership of the Ministry of Foreign Affairs'. This statement was made by Russian President Vladimir Putin on 5 July at a press conference after negotiations with Hungarian Prime Minister Viktor Orban.
Label: Guardian, Peacemaker

B Confusion Matrix