# gowithnlp at SemEval-2025 Task 10: Leveraging Entity-Centric Chain of Thought and Iterative Prompt Refinement for Multi-Label Classification

**Bo Wang[1,3]†, Ruichen Song[1]†, Xiangyu Wang[1],**
**Ge Shi[2]*, Linmei Hu[1]*, Heyan Huang[1,3], Chong Feng[1,3]**
[1]Beijing Institute of Technology, China
[2]Beijing University of Technology, China
[3]Southeast Academy of Information Technology, Beijing Institute of Technology, China

{bwang,src,wwxyy,hulinmei,hhy63,fengchong}@bit.edu.cn shige@bjut.edu.cn * †

## Abstract

This paper presents our system for Subtask 10 of Entity Framing, which focuses on assigning one or more hierarchical roles to named entities in news articles. Our approach iteratively refines prompts and utilizes the Entity-Centric Chain of Thought to complete the task. Specifically, to minimize ambiguity in label definitions, we use the model's predictions as supervisory signals, iteratively refining the category definitions. Furthermore, to minimize the interference of irrelevant information during inference, we incorporate entity-related information into the CoT framework, allowing the model to focus more effectively on entity-centric reasoning. Our system achieved the highest ranking on the leaderboard in the Russian main role classification and the second in English, with an accuracy of 0.8645 and 0.9362, respectively. We discuss the impact of several components of our multilingual classification approach, highlighting their effectiveness.

## 1 Introduction

The task of Entity Framing, as part of the SemEval 2025 campaign, focuses on the automatic identification and classification of roles assigned to named entities (NEs) in news articles(Piskorski et al., 2025). Specifically, it involves determining the roles of protagonists, antagonists, and innocents within the context of news articles related to high-stakes global issues such as the Ukraine-Russia war and climate change. This task is of particular importance as it supports the identification of manipulation attempts and disinformation in media, thereby facilitating a better understanding and analysis of news narratives. The task is multilingual, covering five languages: Bulgarian, English, Hindi, Portuguese, and Russian, and aims to offer valuable insights into how different languages handle entity roles in manipulative content.

Our approach employs a pipeline-based method that automates the optimization of category definitions through iterative refinement (Yang et al., 2023; Xing and Chen, 2024). This process incorporates hard instance analysis to refine category definitions. However, some hard samples may still not be fully covered; we address this by extracting few-shot cases to supplement the category definitions, thereby mitigating misclassifications and resolving ambiguities in category boundaries. Additionally, we leverage entity-centric Chain of Thought (CoT) mechanisms (Wei et al., 2022b) to enhance classification accuracy by focusing on relevant entities and preserving key contextual information, even in lengthy texts. This pipeline process facilitates the continuous improvement of category definitions and prediction refinement through multiple feedback iterations, while also alleviating the challenges posed by an excessive number of categories in multi-label classification tasks.

We ranked second in English and among the top in the Russian language's main role classification subtask. However, challenges remain in handling ambiguous cases and fine-tuning role definitions, which can sometimes impact the precision of role assignments. In our experiments, we further analyzed the effects of prompt iteration optimization, ensemble strategies, and entity-centric CoT on the overall performance.

## 2 Background

The Entity Framing task in SemEval 2025 aims to automatically identify and classify the roles of named entities (NEs) within news articles. The task specifically targets three roles: protagonists, antagonists, and innocents, based on a predefined taxonomy. This is a multi-label, multi-class text-span classification problem, where the goal is to classify NEs within the context of news articles related to high-profile topics. These areas are highly

[*] Co-corresponding Author
[†] Both authors contributed equally to this work.

*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 373–380
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

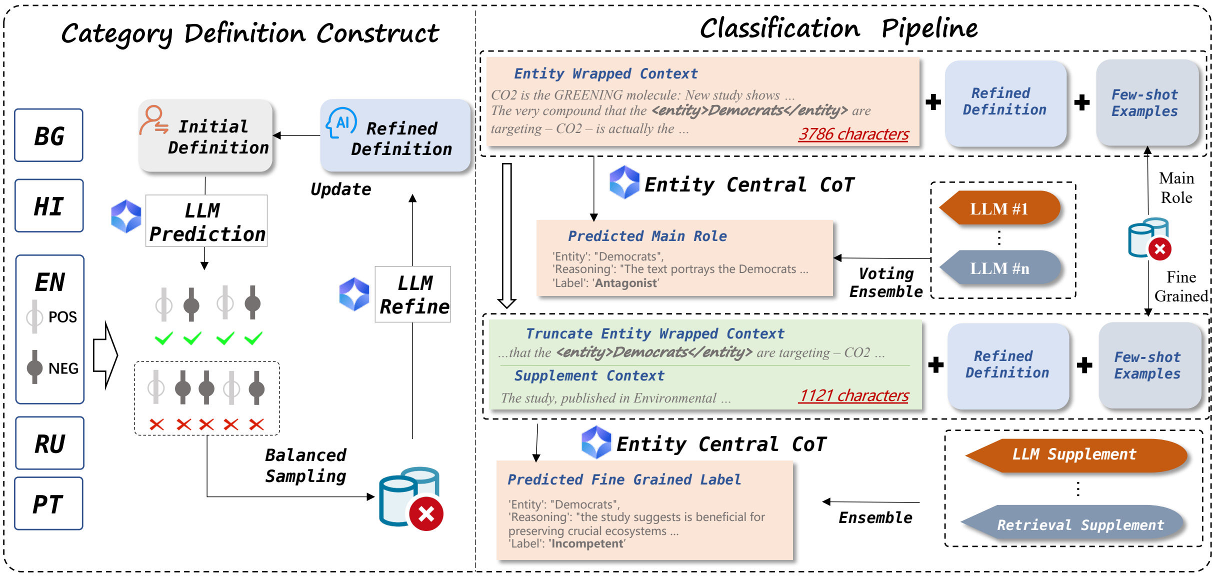

Figure 1: Overview of the Category Definition and Classification Pipeline. The pipeline begins with the construction of initial category definitions using LLM predictions, followed by iterative refinement based on misclassifications. The classification process involves entity context processing with entity-centric Chain of Thought (CoT), combined with few-shot examples and voting ensemble methods.

susceptible to manipulation, misinformation, and disinformation, making the task crucial for understanding how media outlets shape narratives. The dataset provided for this task consists of articles in five languages. The articles, collected between 2022 and 2024, are primarily sourced from alternative media outlets and web portals, many of which have been flagged by fact-checkers for potential misinformation. The detail of datasets is described in Task 10 (Piskorski et al., 2025).

Automatic prompt engineering has gained traction for optimizing prompt design in large language models (LLMs), reducing manual effort. Approaches like Autoprompt (Shin et al., 2020) focus on generating prompts for classification tasks, while PRetrain (Liu et al., 2023) combines prompt learning with pretraining to enhance adaptability. Recent advancements include PromptAgent (Wang et al., 2023b) and Automatic Engineering of Long Prompts (Hsieh et al., 2024). Traditional text classification methods such as RNN (Xie et al., 2020), GCN (Yao et al., 2019), and LLM-based approaches (Lin et al., 2021) laid the foundation. While LLM-generated reasoning is effective for step-by-step problem-solving (Wei et al., 2022a), it faces challenges like unfaithfulness and logical incoherence (Turpin et al., 2023), often due to the influence of explanation tokens. LLMs are increasingly used for classification, both as standalone solutions (Sun et al., 2023) and in multi-task settings (Longpre et al., 2023). Given the characteristics of the Entity Framing task, such as multi-label classification and context-dependent entity roles, we propose a solution that integrates automatic prompt generation and enhanced reasoning structures to improve accuracy and robustness.

## 3 System Overview

To address the challenges posed by the rapid increase in label quantities, which can complicate classification tasks, we employ a pipeline approach. This approach utilizes the results from main role classification as input for fine-grained classification. For both main role and fine-grained tasks, we incorporate iterative prompt refinement to enhance category definitions, followed by inference using an entity-centric CoT framework to improve the model's understanding of entity-based tasks. Subsequent sections will provide a comprehensive overview of our model's key components, including Prompt Construction, Multi-label Classification, and Post-processing.

### 3.1 Category Definition Construction

Although the task provides category definitions based on natural language, directly utilizing the definitions in the prompt for LLM classification tasks may lead to interpretational ambiguities. To enhance the classification performance, we approach this issue from two directions: first, by iteratively refining the prompt to optimize its effectiveness, and second, by selecting few-shot examples to guide the model.

**Category Definition.** For each category $i \in \mathcal{I}$, we begin by selecting a random case for model prediction, thereby partitioning the training set into two subsets: $\mathcal{T}_i$, the correctly classified examples, and $\mathcal{T}_i'$, the incorrectly classified examples. The category definition is then refined by incorporating the original label definition, erroneous examples, and task-specific requirements. The task requirements specify the need to adjust the template to address issues related to input length while improving the model's ability to differentiate between categories. Directly inputting all erroneous examples into the LLM for label redefinition may introduce bias; therefore, a balanced number of false positive and false negative examples are selected from $\mathcal{T}_i'$, ensuring equitable representation across categories. Multiple rounds of prompt refinement are conducted, followed by a comprehensive synthesis of these iterations to finalize the category definitions. The refined category definitions are then used with the updated prompt to predict and iteratively update $\mathcal{T}_i'$, thereby improving the category definitions in subsequent iterations.

**Few-shot Selection.** Due to the inherent limitations of the model's foundational capabilities, there may still be instances of prediction errors. The boundaries of categories, as defined by natural language, cannot always be universally delineated. To address this, we leverage few-shot learning to supplement the categorization process. We aim for a complementary relationship between few-shot examples and prompts. Specifically, we classify the model based on the final category definitions and then select few-shot examples from the resulting set of errors.

The distinction between main roles and fine-grained roles lies in their granularity and complexity. For fine-grained classification and few-shot selection, we similarly prioritize the selection of hard samples. However, due to the small number of samples in each category, when appropriate samples cannot be found within the hard samples, we resort to random selection from the entire category set. To manage the context length effectively, we ensure that all few-shot cases are truncated with an entity-centric approach, thereby preventing the context from becoming excessively long.

## 3.2 Classification Pipeline

Since entities in text often appear with numerous mentions, the attitude or polarity associated with different occurrences of the same entity can vary depending on their position within the text. To address this, we follow the entity-centric generation-based approach (Wang et al., 2023a; Li et al., 2021). Specifically, we wrap the entity mention in the context with special tokens, such as "`<entity>` ENTITY SPAN `</entity>`", to highlight the span of the entity. This technique emphasizes the entity's presence and reinforces its significance within the classification task, thereby improving the model's ability to focus on the correct entity mention during the classification process.

In the task of **main role classification**, the prompt is first designed to clearly define the task, presenting a concise description in a single sentence. The categories are then enumerated in the Description section, with key points highlighted for emphasis. Following this, comprehensive definitions for each category are provided, accompanied by few-shot examples to guide the classification.

To improve the reasoning capacity, we implement an entity-centric CoT approach. Specifically, in addition to the standard reasoning procedure inherent to it, we require the model to first identify the entity to be classified before proceeding with the reasoning step. This modification aims to address the issue of information loss, which frequently arises in long texts due to the model's tendency to lose focus on relevant entities as the text progresses. Finally, the complete text is concatenated and fed into the model. During the classification process, task-specific characteristics introduce variations. We observe that LLMs exhibit biases due to the distribution of training data, which affects the classification of certain entities. To mitigate this, we incorporate the full context, ensuring a more accurate representation of the text's theme and improving entity classification.

We leverage the results above for **fine-grained classification**. In terms of the fine-grained part, the experimental results indicate that longer texts result in a sharp decline in classification performance, even when they contain critical information for the classification task. We truncate the context to enhance the important information. To prevent the loss of distant but relevant information, we incorporate a retrieval-based method to supplement the input. Additionally, during the evaluation process, we experimented with LLM-based summarization to obtain the diversity results.

Table 1: Dataset Statistic, PRO denotes the Protagonist, ANT denotes the Antagonist, INN denotes the Innocent

| LN | Train | | | | Dev |
|---|---|---|---|---|---|
| | PRO | ANT | INN | Total | Total |
| BG | 78 | 425 | 124 | 627 | 31 |
| EN | 130 | 477 | 79 | 686 | 91 |
| PT | 353 | 579 | 319 | 1251 | 280 |
| HI | 1083 | 766 | 482 | 2331 | 116 |
| RU | 222 | 396 | 104 | 722 | 86 |

Table 2: Official SemEval Results on the Test Set

| | Rank | EM | F1 | Acc |
|---|---|---|---|---|
| BG | 4 | 0.4355 | 0.4524 | 0.8306 (#6) |
| EN | 4 | 0.3702 | 0.4160 | 0.9362 (#2) |
| PT | 9 | 0.2694 | 0.3032 | 0.7138 (#6) |
| HI | 7 | 0.3354 | 0.3983 | 0.7152 (#4) |
| RU | 5 | 0.4486 | 0.4671 | 0.8645 (#1) |

Table 3: Results on Development Set.

| Method | EM | F1 | Acc |
|---|---|---|---|
| Ensemble | 0.49450 | 0.53113 | 0.94505 |
| Ours | 0.49450 | 0.53113 | 0.89010 |
| w/o CD | 0.47252 | 0.51648 | 0.86813 |
| w/o FS | 0.45054 | 0.48717 | 0.90109 |
| w/o EC | 0.47252 | 0.50915 | 0.87912 |
| w/o CoT | 0.45054 | 0.48717 | 0.82417 |
| w/o CP | 0.40659 | 0.44322 | - |

### 3.3 Post Process

To improve the robustness of our classification results, we employ an ensemble approach for both the main role and fine-grained classification tasks by utilizing multiple LLMs with different input strategies. These results are then aggregated using two ensemble techniques. The voting mechanism aggregates the predictions by selecting the most frequent classification, while the LLM peer review process involves comparing the outputs from different models and refining the final prediction.

## 4 Experiment

### 4.1 Data Set

As shown in Table 1, the distribution of labels varies across languages. Accuracy(Acc) is applied to evaluate the performance of the main role classification. For fine-grained classification, we employ a combination of Exact Match (EM) and micro F1 score (F1) as evaluation metrics to assess the model's ability to classify roles.

### 4.2 Main Results

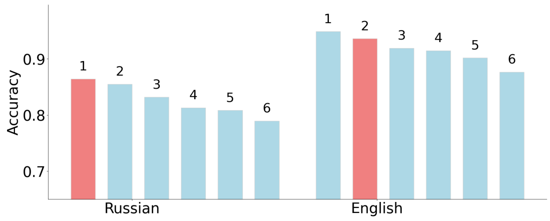

Figure 2: Result on Main Role Classification, our model is highlighted in red, while the top 5 teams are marked in blue excluding ours.

As shown in Figure 2, our system achieved first place in the Russian language's main role classification and second place in English. The results for main role classification demonstrate the effectiveness of our approach. For the fine-grained classification shown in Table 2, We observed that our approach performs the best in Russian and performs the worst in Portuguese. However, the fine-grained results indicate that the model's performance still varies significantly across languages, with certain languages showing weaker performance in specific tasks. These differences highlight the need for further refinement, especially for languages with less optimal results.

### 4.3 Ablation Study

To further analyze the influence of each component, we remove each modular of our method individually and test the influence on the English dataset, as shown in Table 3, where CD denotes the Category Definition, FS denotes the Few-shot Selection, EC denotes the Entity-Centric reasoning, and CP denotes the context compress. To avoid the influence of error propagation, **we use the ground truth of the main role to test the fine-grained classification for the following experiments.** The results demonstrate that the CoT has the most significant impact on model performance. Additionally, when category definition and entity-centric components are removed, there is a substantial drop in model performance. In main role classification, the impact of few-shot selection is relatively minor, possibly due to the smaller number of categories. The phenomenon leads to overall better performance. However, in the fine-grained classification task, hard samples play a more important role in improving classification accuracy.

### 4.4 Influence of Language

Additionally, we explored the influence of prompts and CoT on the performance of the proposed method, and the results are shown in Table 4. Con-

| TEXT | COT | EM | P | R | F1 | Acc |
|---|---|---|---|---|---|---|
| RU | RU | 0.58139 | 0.61627 | 0.59883 | 0.60465 | 0.84883 |
| | EN | 0.58139 | 0.61627 | 0.59883 | 0.60465 | 0.83720 |
| HI | HI | 0.45000 | 0.52857 | 0.48869 | 0.50178 | 0.67142 |
| | EN | 0.46071 | 0.53571 | 0.49761 | 0.51011 | 0.68571 |
| PT | PT | 0.59482 | 0.64655 | 0.62608 | 0.62931 | 0.88793 |
| | EN | 0.65517 | 0.70689 | 0.68103 | 0.68965 | 0.85344 |
| BG | BG | 0.38709 | 0.48387 | 0.43548 | 0.45161 | 0.80645 |
| | EN | 0.35483 | 0.45161 | 0.40322 | 0.41935 | 0.87096 |

Table 4: Performance comparison across different languages for CoT.

ducting chain-of-thought (CoT) reasoning in English generally yields superior performance in the fine-grained classification task, which is relatively complex. For main role classification, reasoning in the same language as the context enhances the accuracy of the results. The observed specificity in Bulgarian can likely be attributed to the small sample size in the development set, which consists of only 31 instances, thus introducing a degree of variability and potential randomness in the outcomes.

## 4.5 Error analysis

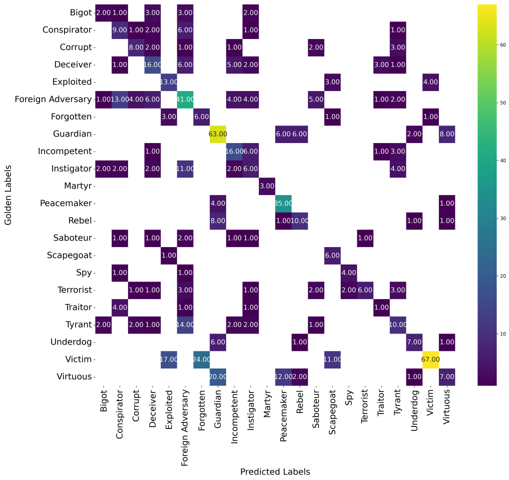

Figure 3: Confusion Matrix

For further analysis, we evaluated the classification performance on each category of the main role and fine-grained label. We observe that severe confusion mainly occurs between the predicted label and golden label, which is shown as Figure 3. The experimental results show that for most cases of most labels, our pipeline can accurately determine the fine-grained role labels.

## 4.6 Post-processing Analysis

To further analyze the applicable scenarios of Voting and LLM peer review ensemble methods, we test the performance of these methods on the English development set. As observed in Table 5, we obtained varying outcomes for the main role and fine-grained classification tasks. For main role classification, we found that the voting ensemble method was more direct and effective. When leveraging LLMs for analysis and reasoning in the main role task, excessive analysis led to hallucinations by the LLM, which in turn degraded the model's ensemble performance. In contrast, for fine-grained classification, which presents a higher level of difficulty, the LLM peer review-based ensemble method proved to be more suitable, better adapting to tasks that require complex reasoning.

Table 5: Results with Different Ensemble Methods.

| Method | EM | F1 | Acc |
|---|---|---|---|
| Voting | 0.49450 | 0.53113 | 0.94505 |
| LLM | 0.50549 | 0.54945 | 0.92307 |

# 5 Outro

This work presents an LLM-based framework to tackle the challenges of entity role classification in news articles, employing iterative prompt refinement and entity-centric chain-of-thought mechanisms. The experiments provide a detailed analysis of the impact of various strategies for employing LLMs in this fine-grained entity understanding task, including language, the use of CoT, et al. These insights offer valuable guidance for future multi-label entity classification tasks based on LLMs.

## 6 Acknowledgement

We sincerely thank the organizers of the shared task and reviewers for their helpful feedback. This work was supported by the National Key Research and Development Program of China No.2022YFC3302101.

## References

Cho-Jui Hsieh, Si Si, Felix Yu, and Inderjit Dhillon. 2024. Automatic engineering of long prompts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10672–10685.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gcn and bert. *arXiv preprint arXiv:2105.05727*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

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.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8990–9005. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Bo Wang, Heyan Huang, Xiaochi Wei, Ge Shi, Xiao Liu, Chong Feng, Tong Zhou, Shuaiqiang Wang, and Dawei Yin. 2023a. Boosting event extraction with denoised structure-to-text augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11267–11281, Toronto, Canada. Association for Computational Linguistics.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023b. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Xintao Xing and Peng Chen. 2024. Entity extraction of key elements in 110 police reports based on large language models. *Applied Sciences*, 14(17):7819.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

## A Prompt Construction

The overall structure of our classification prompt is provided, where the blue sections represent entries obtained either from predefined sources or from the training set.

**Prompt For Entity Classification**

**# Task Description:**
...
**# Category Definitions:**
## Category #1
...
**# Few-Shot Example:**
## Example #1
...
**# Requirements:**
...
**# Output Format:**
Entity: ...,
Reason: ...,
Label: ...
**# Wraped Context:**
...

## B Implement Details

In the prompt iteration optimization, each round involves using 2 false positive and 2 false negative examples. Additionally, we integrate 5 different definitions to form the final category definition. For the classification task, we utilize models including GLM-4-Plus[1], Claude 3.5 Sonnet[2], and GPT-4o[3] to perform the task. For the main role classification, we integrate X models from different foundational LLMs and methods. We retain the models that perform best within 5% of the highest score in the development set, ensuring robust integration. In this case, we incorporate no fewer than three models for each language. In the fine-grained role classification, we integrate systems that include a variety of few-shot selection and compression methods. To capture the diversity of approaches, at least five distinct results are considered for this task. As with the main role classification, the final output from the ensemble is empirically adjusted based on empirical adjustments to optimize overall accuracy. The GLM-4-Plus may refuse to provide a response during prediction due to alignment, we substitute the response with GPT-4o if the model fails to answer after 5 rounds.

[1] https://open.bigmodel.cn/dev/api/normal-model/glm-4
[2] https://www.anthropic.com/news/claude-3-5-sonnet
[3] https://openai.com/

## C Results of Main Role

In addition to English and Russian, we present the ranking results for the other three languages, as shown in Figure 4.

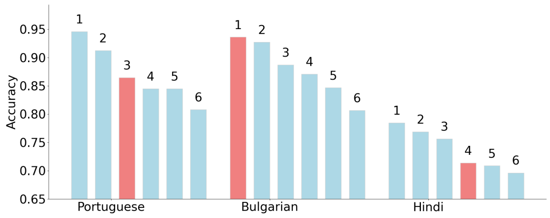

Figure 4: Result on Main Role Classification, experiment setting is same with Figure 2.

## D Error Analysis of Ensemble Model

Multi-label instances are rare in the development dataset, so we split the multi-label instances into multiple individual entries to plot the confusion matrix. The confusion matrix of the ensemble model is shown in Figure 5. The experimental results show that for most cases of most categories, our pipeline can accurately determine the fine-grained role labels.

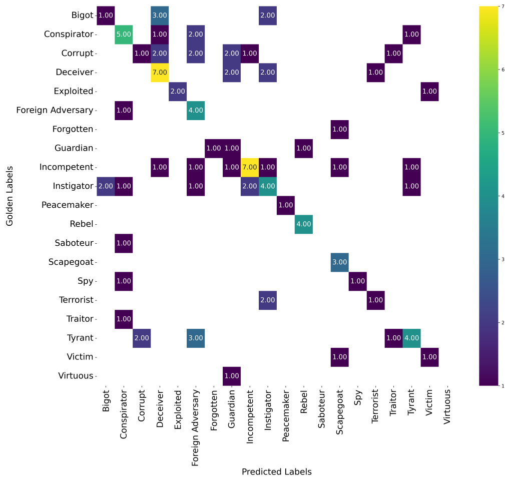

Figure 5: Confusion Matrix of Ensemble Pipeline

## E Analysis of Compression Methods

We further investigated the influence of different compression methods. Excessively long input text may lead to a loss of detailed information within LLMs. To mitigate the adverse effects of long text in fine-grained classification, we designed several text compression strategies. PLAIN uses the whole original context as the input of classification. TRUNCATION use a fixed number of sentences

| Method | EM | F1 |
|---|---|---|
| Plain | 0.41758 | 0.45421 |
| Truncate | 0.48351 | 0.52014 |
| Retrieve | 0.46153 | 0.50549 |
| Character | 0.47252 | 0.50183 |
| LLM | 0.41758 | 0.45421 |

Table 6: Performance comparison across different methods for text compression.

which will retained both above and below the sentence containing the target entity. RETRIEVAL retrieve additional contextually relevant sentences as supplement context. CHARACTER uses the character number to truncate the context instead of the number of sentences. LLM generates a summary while preserving the original entity information. Experimental results shown in Table 6 demonstrate that the performance of text input is generally better after compression.