DUTIR at SemEval-2025 Task 10: A Large Language Model-based Approach for Entity Framing in Online News

Tengxiao Lv, Juntao Li, Chao Liu, Yiyang Kang, Ling Luo*, Yuanyuan Sun, Hongfei Lin
School of Computer Science and Technology, Dalian University of Technology, China
{tengxiaolv, juntaoli, liuchao2464687308, kiang0920}@mail.dlut.edu.cn,
{lingluo, syuan, hflin}@dlut.edu.cn
***Correspondence:** lingluo@dlut.edu.cn

Abstract

This paper presents our system designed for Subtask 1 of SemEval-2025 Task 10, which focuses on multilingual entity framing in news articles. Given the complexity of the task, which involves multi-label, multi-class classification across five languages, we propose an approach based on large language models (LLMs). This approach combines multilingual text translation, data augmentation, multi-model fine-tuning and ensemble classification. First, we translated texts into English to unify the datasets, followed by synonym-based augmentation to address class imbalances. We then fine-tuned multiple LLMs using the augmented dataset. Finally, a cutting-edge LLM was applied to aggregate model predictions for ensemble classification, ensuring robust and accurate classifications. Our system demonstrated promising results, achieving top positions in three languages (English, Portuguese and Russian) and second place in Bulgarian.

1 Introduction

With the increasing prevalence of the internet, people can easily access diverse information, which has also facilitated the propagation of misinformation more readily compared to traditional media. Public perceptions of events are often influenced by these harmful false narratives and propaganda, particularly regarding major crisis incidents. Consequently, misinformation identification has become crucial, prompting growing research (Orbach et al., 2021; Sharma et al., 2023) to analyze and categorize entities in textual information.

The SemEval-2025 Task 10 (Piskorski et al., 2025; Stefanovitch et al., 2025) focuses on multilingual representation and narrative extraction from online news, aiming to advance research and development of novel analytical capabilities to support end-users in analyzing news ecosystems and identifying characteristics of manipulation attempts. The task organizers construct a dataset (Mahmoud et al., 2025) comprising 1,378 news articles focusing on the Ukraine-Russia war and climate change, with role annotations applied to over 5,800 entities. This task comprises three subtasks and we participate in Subtask 1 (Entity Framing). Specifically, given a news article and a list of Named Entity (NE) mentions within it, the objective is to assign one or multiple roles to each mention using a predefined refined role taxonomy. This taxonomy encompasses three primary role types: protagonists, antagonists and innocent, forming a multi-label multi-class text span classification task.

Entity Framing subtask presents two main challenges. First, as a multilingual task involving five languages (Bulgarian, English, Hindi, European Portuguese and Russian), traditional methods exhibit limited modeling capabilities for large-scale, complex multilingual tasks, particularly in handling long sequences and intricate semantics. The emergence of LLMs (Zhao et al., 2023; Matarazzo and Torlone, 2025) addresses this challenge effectively through their robust generalization capabilities and multilingual data integration. Second, this subtask requires multi-label multi-class classification where each entity must be assigned to one of three primary roles, with further granularity in secondary subcategories. For instance, the protagonist category includes finer-grained roles such as Guardian, Martyr, Peacemaker, Rebel, Underdog and Virtuous. This hierarchy demands enhanced semantic comprehension and classification capabilities from models.

To address these challenges, we implemented an LLM-based pipeline. Initially, we translated all non-English data into English for unified processing and constructed a task-specific dataset. We then fine-tuned multiple foundational large language models (GLM et al., 2024; Yang et al., 2024; Dubey et al., 2024) on this dataset. Subsequently, we employed a state-of-the-art LLM by designed

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 174–179

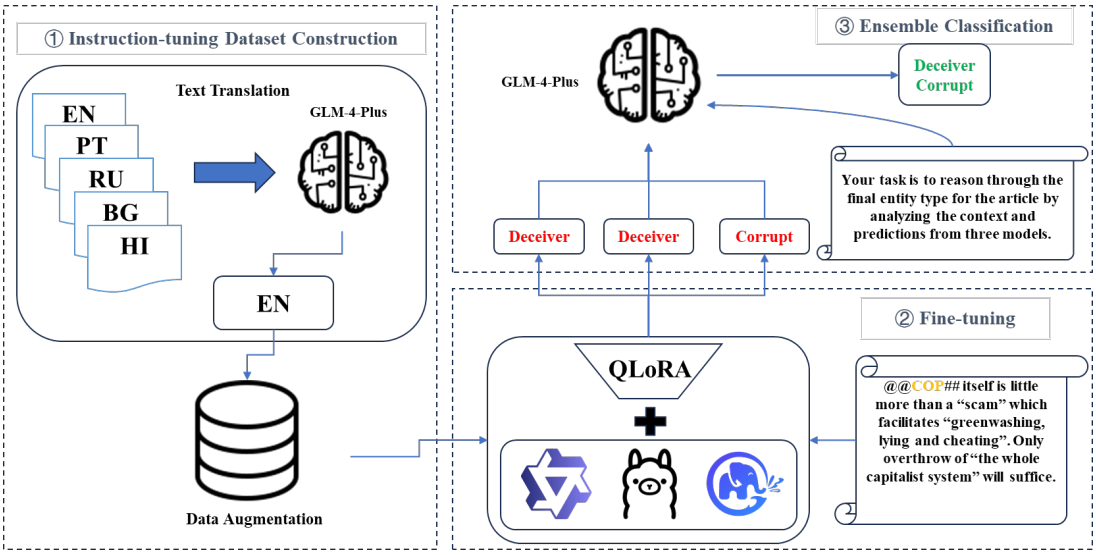

Figure 1: The overall architecture diagram of our system. The "COP" highlighted in yellow is the entity to be classified, the one marked in red indicates the incorrect classification result, and the one marked in green indicates the correct classification result.

ensemble prompts to aggregate decisions from multiple LLMs, generating final classification results. Our proposed method demonstrated strong performance by achieving first place in three out of the five languages evaluated in Subtask 1.

2 Background

Named Entity Recognition (NER) has long been a key research direction in the field of Natural Language Processing (NLP) (Xu et al., 2024). NER refers to the task of identifying entities with specific meanings in text, such as person names, locations and others, and annotating them accordingly. Essentially, it is a sequence labeling task aimed at classifying each word or phrase in a text as belonging to a specific named entity category or not belonging to any named entity category.

In recent years, with the emergence of an increasing number of open-source large models (Touvron et al., 2023; Liu et al., 2024) and the introduction of various fine-tuning techniques (Hu et al., 2021; Dettmers et al., 2024), LLMs have achieved significant progress in NER tasks (Luo et al., 2024). In this study (Naguib et al., 2024), they collect and use 14 NER datasets covering English, French, and Spanish, and compare the performance of generative LLMs with few-shot prompts and traditional masked-based models in both general and clinical domains. GPT-NER (Wang et al., 2023) cleverly transforms the traditional NER sequence labeling task into a generation task that is easier for LLMs to handle, using special tokens to mark the entities to be extracted. Additionally, it constructs few-shot prompt words by retrieving semantically similar examples from the input via KNN, effectively bridging the gap between the NER task and LLMs.

3 System Overview

As shown in Figure 1, the overall structure of our system includes the following key components: Instruction-tuning dataset construction based on multilingual text translation and data augmentation, multi-model fine-tuning based on QLoRA, and ensemble classification based on GLM-4-Plus.

3.1 Instruction-tuning Dataset Construction

Due to the superior model capabilities and abundant data annotation resources of LLMs in English, coupled with the relatively smaller scale of data annotation in other languages, translating multilingual texts into English can effectively expand the data scale, thereby simplifying classification tasks and reducing the complexity of processing multilingual texts. Based on this, the first step of the system is to translate texts in various languages into English. This process is achieved by calling the API of GLM-4-Plus, which supports multilingual translation. The translated texts are then uniformly consolidated into a corpus for subsequent processing.

However, there is an issue of imbalance in the types of entities in the dataset. To mitigate this problem, we further employed data augmentation

###Instruction:
Below is a news article. For each named entity marked with @@## in the article, classify its role based on a predefined taxonomy of fine-grained roles. Each entity can have one or more roles.

Input:
Bill Gates Says He Is 'The Solution' To Climate Change So It's OK To Own Four Private Jets \n\n@@Bill Gates## has the right to fly around the world on private jets… Predefined taxonomy:\n– Guardian: …\n– Martyr: …\n– Peacemaker:…

output:
Named entities marked with @@## in the article: Bill Gates; Based on the taxonomy, the classification is: Deceiver, Corrupt.

Figure 2: Example of the prompt from our training set. The "Bill Gates" highlighted in yellow is the entity to be classified, and the one marked in green indicates the correct classification result.

####System
Your task is to reason through the final entity type for the article by analyzing the context and predictions from three models. The final type may have one or more entities, separated by a "、".
Please follow these steps:
1.Analyze the sentence: Understand the semantics of the article and the context in which the entity appears.
2.Evaluate the predictions: Review each model's prediction to determine whether it is reasonable.
3.Resolve contradictions: If there is disagreement in predictions, decide which one better fits the context.
4.Make a final decision: Arrive at the final conclusion by weighting the evidence or semantic fit.
Please show the thought process clearly, and mark the final result in [square brackets].
###Input
Article: @@COP## itself is little more than a "scam" which facilitates "greenwashing, lying and cheating". Only overthrow of "the whole capitalist system" will suffice.
Entity to be marked: COP
Model predictions: Qwen2.5: Deceiver、llama3.1: Deceiver、GLM4: Corrupt

Figure 3: Example of the Chain-of-Thought Prompt for LLM-based Ensemble Learning.

	Train	Dev	Test	AVG Length
EN	686	91	235	1646
PT	1251	116	297	2269
RU	722	86	214	2433
BG	627	31	124	3239
HI	2331	280	316	8190
DA	700	-	-	1521
Total	6317	604	1186	4418

Table 1: Statistics of dataset sizes. DA represents the dataset obtained through data augmentation, and AVG Length refers to the average length of the training set.

strategies, especially synonym replacement for underrepresented entity types, to enhance their representation in the training set (Dai and Adel, 2020). Specifically, we utilized WordNet as the synonym dictionary and applied probabilistic replacement to 30% of non-entity words in sentences containing rare entity types that occur less than 30 times. The augmented data, combined with the original dataset, forms a more balanced training set. This strategy not only effectively addresses the issue of class imbalance but also lays a solid foundation for subsequent model training while ensuring the consistency of input formats.

3.2 Multi-model Fine-tuning

During the model training phase, we employ QLoRA technology to fine-tune multiple base models, which include Qwen2.5, Llama3.1, and GLM4. The choice of QLoRA is due to its ability to perform efficient fine-tuning operations with limited computational resources, making it particularly suitable for LLMs.

The fine-tuning dataset's prompt is shown in Figure 2. Specifically, each model's fine-tuning prompt includes a task description, candidate entity types and their definitions, and special symbols to mark the positions of entities in the text. For example, in combination with an automatic labeling program, the symbol "@@" is used to mark the start of an entity, and the symbol "##" is used to mark the end of an entity.. This design helps the model better understand the task by first locating entities in the text through prompts and then proceeding with classification. In the fine-tuning process, each base model is trained on the augmented dataset, with the goal of optimizing the model's parameters to minimize classification loss.

3.3 Ensemble Classification

After fine-tuning each base model, we designed an LLMs prompt-based ensemble learning, which employed a Chain-of-Thought approach to guide the LLMs in analyzing the classification results of entities based on multiple models fine-tuned with QLoRA, thereby obtaining the final results. The specific prompt is provided in Figure 3. Each fine-tuned model (Qwen2.5, Llama3.1, and GLM4) generates a classification result for a given entity. These results are then passed to the GLM-4-Plus model, which acts as a meta-classifier to conduct a comprehensive analysis of all the models' prediction outcomes and ultimately make a decision.

As shown in Figure 1, for the sentence

	EN			PT			RU			BG			HI		
	EM	**F1**	**Rank**	**EM**	**F1**	**Rank**	**EM**	**F1**	**Rank**	**EM**	**F1**	**Rank**	**EM**	**F1**	**Rank**
PATeam	38.30	44.53	2	49.16	53.97	2	44.39	49.33	6	**51.61**	53.54	1	26.90	32.05	11
DEMON	37.45	42.08	3	36.70	41.36	6	46.73	49.66	4	45.97	47.01	3	40.19	47.56	4
QUST	32.77	37.98	7	45.79	49.28	3	51.40	54.75	2	38.71	38.74	6	**46.84**	**53.85**	1
TartanTritons	35.74	10.78	5	33.33	17.42	8	47.20	15.80	3	41.13	9.52	5	44.62	17.33	2
BERTastic	25.11	29.60	11	41.75	45.48	4	46.73	48.98	4	35.48	36.51	7	43.99	51.57	3
Baseline	3.83	4.40	27	4.71	4.84	15	5.14	5.90	15	4.03	3.97	14	5.70	7.16	15
DUTIR(our)	**41.28**	**45.42**	1	**59.26**	**63.72**	1	**56.54**	**60.36**	1	50.81	**54.96**	2	29.43	34.10	8

Table 2: Leaderboard of the test set. The table presents the leaderboard results for the test set, with the Exact Match Ratio (EM) and micro F1 score displayed as percentages. The best results for each metric are highlighted in bold. Additionally, we list the top three teams in any language, with the possibility of ties in ranking, as well as baseline results for comparison. The ranking is based on the Exact Match Ratio (EM).

"@@COP## itself is little more than a 'scam' which facilitates 'greenwashing, lying and cheating'. Only overthrow of 'the whole capitalist system' will suffice. ", where "COP" is the entity to be categorized, Qwen2.5 classifies it as "Deceiver", Llama3.1 classifies it as "Deceiver", and GLM4 classifies it as "Corrupt". Based on the prediction results of each fine-tuned model, and considering their performance and reliability in specific tasks, GLM-4-Plus makes the final entity classification decision. By adopting this ensemble method, we can effectively enhance the accuracy and robustness of classification, especially when dealing with complex or diverse entity types. Ensemble learning fully leverages the strengths of each base model to achieve more precise classification results.

4 Experimental Setup

The dataset originates from Subtask 1 of Task 10 in SemEval 2025, comprising news articles in plain text format across five languages: English (EN), Portuguese (PT), Russian (RU), Bulgarian (BG) and Hindi (HI).

During the experimental phase, we generated individual data records for each entity mention, with the statistical summary presented in Table 1. The limited number of available articles in each language undoubtedly increased the difficulty for LLMs to learn effectively. Additionally, the dataset exhibited significant class imbalance, further intensifying the challenge of the task.

To address these challenges, we employed strategies of data translation and augmentation. Specifically, we leveraged LLMs to translate all articles from different languages into English, resulting in 5,617 data records. Building on this, we conducted data augmentation operations such as synonym replacement for underrepresented categories, adding additional 700 data records. Ultimately, all datasets were merged to form a complete dataset containing 6,317 training data records.

For task evaluation, the official assessment utilized multiple metrics to comprehensively measure model performance, including Exact Match Ratio, micro precision (micro P), micro recall (micro R), micro F1 score (micro F1), and accuracy for the main role. Among these, the Exact Match Ratio was the primary evaluation metric.

During the training process, a batch size of 4 was used, the learning rate was set to 1e-4, and a maximum truncation length of 4096 was set to accommodate text inputs of varying lengths. Additionally, the AdamW optimizer was selected to further enhance the model's training efficiency and generalization ability. All experiments were conducted on a single NVIDIA L40 GPU.

5 Results

5.1 Final Submission

The detailed information of the test set leaderboard is shown in Table 2. Ranked by exact match rate, our proposed method achieved first place in three out of the five languages covered. In addition, our system achieved the highest micro F1 score in four languages, fully demonstrating its effectiveness and adaptability. However, its performance on Hindi was unexpectedly unsatisfactory.

After analysis, we believe that this result may be closely related to the length characteristics of Hindi articles. The average lengths of datasets for different languages are shown in Table 1. Compared to other languages, Hindi articles are generally longer. Due to hardware limitations, we were unable to set a longer token length.

5.2 Ablation Study

To comprehensively verify the key contributions of each component in the system to overall perfor-

	EN		RU	
	EM	Δ	**EM**	Δ
Our System	**41.28**	-	**56.54**	-
w/o TT	38.30	-2.98	54.67	-1.87
w/o DA	40.43	-0.85	55.14	-1.40
w/o EC	37.87	-3.41	53.27	-3.27

Table 3: The results of the ablation studies on the EN and RU test sets.

mance, we designed the following ablation experiments: we evaluated the system's performance under conditions where multilingual text translation was excluded (marked as w/o TT), data augmentation was not implemented (marked as w/o DA), and ensemble classification was not used (marked as w/o EC). The results of the ablation study are shown in Table 3.

- By introducing LLMs for multilingual text translation, we successfully integrated datasets from multiple languages, providing the model with more comprehensive and in-depth learning materials. This significantly enhanced the model's learning effectiveness and generalization ability.
- For categories with low representation in the dataset, we employed data augmentation strategies, which effectively alleviated the issue of data imbalance. This improved the model's accuracy and robustness when dealing with imbalanced datasets.
- Furthermore, by leveraging high-performance LLMs, we fine-tuned different base models and performed ensemble classification on their classification results. This innovative approach not only further improved the overall performance of the system but also made the system's output more stable and reliable, demonstrating the unique advantages of ensemble learning in enhancing model performance.

6 Conclusion

This paper presents the system we designed for Subtask 1 of SemEval-2025 Task 10. We propose a multilingual text processing framework that combines multilingual translation with data augmentation, QLoRA-based multi-model fine-tuning, and GLM-4-Plus-based ensemble classification. By using GLM-4-Plus to translate multilingual texts into English, we enhance data diversity and quantity. Data augmentation effectively improves the model's performance on imbalanced datasets. QLoRA fine-tuning optimizes the model and reduces classification loss. GLM-4-Plus, as a meta-classifier, further enhances system performance. Our system achieved first place in three languages (English, Portuguese and Russian). In the future, we will focus on improving long-text processing and optimizing LLMs fine-tuning techniques.

7 Acknowledgments

The authors thank the organizers of the SemEval-2025 Task 10. This research was supported by the grants from the National Natural Science Foundation of China (No. 62302076, 62276043).

References

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, page ocae037.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, et al. 2025. Entity framing and role portrayal in the news. *arXiv preprint arXiv:2502.14718*.

Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040*.

Marco Naguib, Xavier Tannier, and Aurelie Neveol. 2024. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852.

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9154–9173.

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.

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? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2149–2163.

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*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.