The Potential of Large Language Models in Assisting Data Augmentation for East Asian Digital Humanities

Stanford CS224N Custom Project

Fengyi Lin
Department of East Asian Languages and Cultures
Stanford University
linfy@stanford.edu

Abstract

The advent of digital humanities prior to machine learning relied heavily on pre-designed, inflexible algorithms such as word counting and topic modeling. While these approaches could reveal certain intrinsic textual structures, they fell short in addressing complex humanitarian questions, such as the association between women and domesticity in late imperial classical Chinese literature. The predictive nature of machine learning models, particularly neural-based ones, offers a solution to these limitations by allowing for the assignment of customized labels to data points, thus expanding the capacity of digital humanities. Despite these advancements, East Asian Digital Humanities (EADH) remains underdeveloped, primarily due to the lack of manually labeled data. Large Language Models (LLMs) have shown promise in generating labels for unlabeled datasets, but their effectiveness in producing high-quality data for EADH is still under investigation.

This project explores the practicality of using LLMs for data augmentation in East Asian humanities, specifically focusing on author attribution tasks for classical Chinese poetry. The study evaluates the performance of Meta Llama-3 Chat (70B) compared to a fine-tuned BERT-based model, GuwenBERT, and a trivial baseline. The results indicate that while LLMs can generate labels, their performance is currently inferior to fine-tuned models, suggesting that extensive application of LLMs in EADH might be premature. Instead, fine-tuning pre-trained models like GuwenBERT shows more promise in this field. The study also highlights the challenges LLMs face in handling classical Chinese texts, such as potential confusion between classical and modern Chinese. The findings underscore the need for further research into more suitable LLMs for EADH and suggest future avenues for enhancing data augmentation techniques in this underexplored area.

1 Key Information to Include

- Mentor: Moussa Doumbouya
- External Collaborators (if any): Xi He, Nan Xu, and Hanqi Zhou from the Department of East Asian Languages and Cultures who participated in the human evaluation.
- Sharing project: N/A
- Contribution: All work was done by Fengyi Lin

2 Introduction

Digital Humanities before the age of machine learning relied heavily on pre-designed, inflexible algorithms. These approaches ranged from word counting to topic modeling based on latent Dirichlet allocation (LDA) (Novikova and Novikov, 2024). These approaches focused on revealing some intrinsic textual structures of the texts, such as the distribution of certain words and sets of words that appear near each other in the texts. While they could provide some concrete evidence for interpretations, these approaches were not good at solving more complex humanitarian questions, such as investigating the association between women and domesticity in late imperial classical Chinese literature.

The predictive nature of machine learning models, especially neural-based ones, makes them good at handling nuanced tasks in humanities. Unlike pre-designed algorithms, in a machine learning context, humans can assign customized labels to each data point with specific needs in mind. For example, to examine the relationship between women and domesticity, one can train a model to do two tasks: determine the presence of women in a passage and whether the passage takes place in a domestic space. Consequently, this approach expands the capacity of Digital Humanities and has great potential to deepen human understanding of literary works.

However, compared to its counterpart in the English language, East Asian Digital Humanities (EADH) is a field that lacks development (Schonebaum, 2022). One reason is the insufficient amount of manually labeled data. Fortunately, Large Language Models have shown great potential to produce labels for unlabeled datasets (Zhou et al., 2024). However, some questions remain to be answered. To what extent can LLMs generate high-quality labeled datasets? Are LLMs superior to humans or still requiring improvements? Given that most LLMs are trained on English corpora, what is the performance of LLMs in the specific context of East Asian literary studies?

Hence, in this project, we investigate the practicality of using LLMs for data augmentation for East Asian humanities, particularly focusing on author attribution tasks. We compare the performance of LLMs with that of a fine-tuned BERT-based model and a trivial baseline.

3 Related Work

3.1 Pre-LLM Data Augmentation

Prior to the advent of LLMs, data augmentation (DA) in NLP primarily relied on traditional techniques. Feng et al. (2021) conducted a comprehensive survey on DA methods in NLP, detailing various techniques employed to enhance datasets. However, these techniques often required a deep linguistic understanding of the tasks at hand. Hedderich et al. (2020) provided a notable survey focusing on data augmentation for low-resource scenarios. They highlighted that these traditional approaches did not fully leverage the potential of LLMs.

3.2 LLM-assisted Data Augmentation

The advent of LLMs has significantly transformed data augmentation in NLP (Zhou et al., 2024). LLMs' nuanced understanding of language has addressed several limitations of previous methods, such as the generation of synthetic data of poor quality, noise, and adverse effects on model performance (Møller et al., 2024)(Li et al., 2023b)(Ye et al., 2024). Ding et al. (2024) surveyed the impact of LLMs on data augmentation, particularly in the context of NLP tasks. Their survey presented a taxonomy of LLM-assisted DA, categorizing it into four types of tasks: data creation, labeling, reformation, and co-annotation. Additionally, they identified four key challenges: data contamination, controllable DA, culture-aware DA, and multimodal DA. Notable applications of LLM-assisted DA include Jahan et al. (2024) work on hate speech detection and Zhang et al. (2024a) research on few-shot text classification.

3.3 Data Augmentation for East Asian Languages

Despite the advancements in LLM-assisted DA, there are relatively few studies focusing on East Asian languages. The work of Zhang et al. (2024b) on LLM-assisted data augmentation for Chinese dependency parsing is one example. Another example is the study of Yao et al. (2024) on Chinese name entity recognition. However, the application of LLM-assisted DA to other tasks and languages within this region remains underexplored.

3.4 Contribution

This work aims to fill this gap by applying LLM-assisted data augmentation to a new task—text classification for classical Chinese poetry. This application extends the current understanding and utility of LLMs in enhancing datasets for underrepresented languages and tasks.

4 Approach

4.1 Task Description

Due to the lack of manually labeled datasets in EADH, We chose a task with pre-determined labels. Authorship attribution for classical Chinese poetry is a good candidate to benchmark the quality of LLM-generated labels. First, the authorship of most poems is already known. Second, authorship attribution is a difficult task as it requires an understanding of each author's writing style, which is highly nuanced. If LLMs perform well in this task, this approach is likely to generalize well to simpler tasks. Third, classical Chinese is a low-resource language. Successful completion of this task implies that the approach might also apply to other low-resource languages, not restricted to Asian languages. We have not found any existing scholarship that examines the quality of LLM-generated datasets in **classical Chinese** yet.

This project hypothesizes that each author possesses a unique quality which is invariant across poems of different genres and different topics. Hence, we could differentiate poets by just examining the poems written by him.

This project includes two authorship attribution tasks, both being five-class classification tasks. We have chosen 10 authors from the Tang dynasty of China (618-907), 5 with the most poems preserved and 5 with the most distinctive styles (subjective judgment). That is, we choose the tasks with respect to Tang poets from two dimensions. The first quantitative dimension is the number of works. By choosing the poets with the most works, we minimize the influence of insufficient data. The second qualitative dimension is the distinctiveness of authorness. This dimension is designed in accordance with the authorness hypothesis.

4.2 Baselines

We have two baselines. The first baseline is a trivial classifier based on the proportion of the number of each author's poems in the corpus. That is, the classifier will predict the author of every poem to be the author with the most poems in the task.

The second baseline is the prediction of fine-tuned GuwenBERT, a model adjusted specifically for classical Chinese NLP tasks (Ethan-yt, 2020). Based on RoBERTa, GuwenBERT is further trained on the Daizhige dataset (garychowcmu, 2019), the most comprehensive dataset for classical Chinese texts at this moment. GuwenBERT outperforms all the other BERT-based models in WYWBenchmark, a benchmark designed for classical Chinese NLP tasks Zhou et al. (2023).

We expect that the performance of LLMs will be better than the first baseline and worse than the second baseline. The code for the second baseline is written by me (the first baseline does not require coding).

4.3 LLMs

4.3.1 The Model Chosen

The LLM chosen for this project is Meta Llama-3 Chat (70B) (Touvron et al., 2023). We chose this model for several reasons. First, the large number of parameters indicates the model's capacity to handle more complex tasks. So, it has a higher chance of performing well in the task of authorship attribution of classical Chinese poems. Second, the model is multilingual. Although it might not be pre-trained on classical Chinese, it possesses some knowledge of modern Chinese, which is related to classical Chinese.

4.3.2 Other Models not Chosen

Before settling on this model, We explored the performances of some other LLMs available on the Together AI platform. The reasons that other LLMs were not used are listed below:

- Some LLMs output several sentences instead of only the name of the predicted author, despite prompt specifications. The models not chosen for this reason include Mixtral-8x7B Instruct v0.1 and Gemma Instruct 7B.
- Some LLMs do not recognize classical Chinese very well and output only empty strings. Llama-2-7B-32K-Instruct falls in this category.
- Some LLMs act exactly like a trivial classifier and output only one poet's name (or only two or three poets' names in some cases) for all input poems. Meta Llama-3 Chat (8B) belongs to this category.
- Some LLMs not only give the Chinese name of the predicted poet but also include the romanization of Chinese characters (pinyin) of the poet in the output. However, the output Chinese characters and pinyin do not match. Meta Llama-3 Chat (8B) belongs to this category. For example, the model acts as a trivial classifier and outputs " 李商隐 (Li Bai)" for all input poems. "Li Bai" here actually corresponds to the poet 李白 who is outside the range of possible poets.

4.3.3 Prompting

For each task, We will use three different prompts (Table 1) with the same meaning and choose the poet with the most votes to increase the stability of the generated results (Li et al., 2023a). For each prompt, We will use one-shot and five-shot in-context training. The one-shot prompting ensures the output of the LLM is in the desired format. In the five-shot prompting, the model is fed with five poems from each poet with correct labels. We expect the LLM would perform better with five-shot prompting since it is supposed to partially "learn" the styles of the five poets from the representative poems. The code for implementing the models is written by me.

The three prompts used in the project are listed below:

Prompt	**Type**
Please determine which one of the five poets is the author of the given poem. Output the name of the predicted poet only. *Authors:* *Poem:*	Instruction
Please identify which of the five poets authored the provided poem. Output the name of the predicted poet only. *Authors:* *Poem:*	Paraphrase
Which one of the five poets is the author of the given poem? Output the name of the predicted poet only. *Authors:* *Poem:*	Question Answering

Table 1: The three types of prompts used in the project

In addition to the three types of prompts listed above, We also tried two other types of prompts in the exploration stage of the project. However, the two types of prompts significantly influenced the output of the LLMs and were thus not used in the actual implementation. For the "Sequence Swapping" type of prompts, some LLMs ignored the instructions and started to output sentences instead of the name of the predicted poet. For the "Confirmation Bias" type of prompts, all LLMs tested in the scope of this project simply output the poet suggested in the prompts.

Prompt	**Type**
Authors: *Poem:* Please determine which one of the five poets is the author of the given poem. Output the name of the predicted poet only.	Sequence Swapping
Please determine which one of the five poets is the author of the given poem. We think the author is (*Some Poet among the Five Poets*). Output the name of the predicted poet only. *Authors:* *Poem:*	Confirmation Bias

Table 2: The two types of prompts not used in the project

5 Experiments

5.1 Data

The corpus used in this project is *Complete Tang Poems* (1705), a collection of Tang poems compiled in the Qing dynasty. The corpus contains over 50,000 poems from over 2,200 poets. The corpus is obtained from a GitHub repository (jackeyGao, 2024). The poets chosen for the four tasks are shown in Table 3.

Task 1	Task 2
Bai Juyi 白居易 2914	Li Shangyin 李商隱 606
Du Fu 杜甫 1472	Wen Tingyun 温庭筠 432
Li Bai 李白 1159	Cen Shen 岑參 408
Yuan Zhen 元稹 837	Jia Dao 賈島 407
Liu Yuxi 劉禹錫 867	Wang Wei 王維 399

Table 3: Poets selected for the two tasks and the number of their poems

In terms of data pre-processing, all poems are cleaned such that (1) punctuation, titles, comments, and characters unrecognizable by UTF-8 encoding are removed and (2) the length is between 20 and 500 Chinese characters. The following are two poems and their labels from task 1 and task 2, respectively (each row is a poem):

白玉誰家郎回車渡天津看花東上陌驚動洛陽人 Expected Output: " 李白"
那知芳岁晚坐见寒叶堕吾不如腐草翻飞作萤火 Expected Output: " 岑參"

5.2 Evaluation Method

To comprehensively examine the quality of LLM-generated labels compared to the two baselines, We will use accuracy, F1 score, precision, and recall as evaluation metrics. The primary metrics of this project are accuracy and F1 score. We use accuracy because it is more comparable with human evaluation results (see the "Human Evaluation" section); we use F1 score because the data for task 1 is slightly imbalanced.

5.3 Experimental Details

5.3.1 GuwenBERT Baseline: Without Fine-tuning

For text classification without fine-tuning, We use the model to tokenize the poems and turn them into text embeddings. We use the base version of GuwenBERT `model_name="ethanyt/guwenbert-base"`. We then use a softmax function to do the classification tasks. We use the conventional 60:20:20 training/validation/test split. We use L2 regularization to avoid overfitting. The value of λ is determined by the validation set. We try $\lambda = 0.01, 0.1, 1, 10, 100$. Other parameters are as default in the `sklearn` library.

5.3.2 GuwenBERT Baseline: Fine-tuned

When fine-tuning the model, We use an 80:20 training-test split. We use the base version of GuwenBERT `model_name="ethanyt/guwenbert-base"`. Some of the training arguments include: `learning_rate=2e-5`, `num_train_epochs=3`, and `bf16=True`. For regularization, We set the dropout rate to 0.3. Other parameters are as default in the `transformers` library. The loss curves for tasks 1 and 2 are shown below.

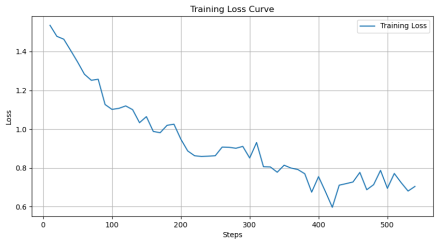

Figure 1: Training Loss Curve of Fine-tuning GuwenBERT for Task 1

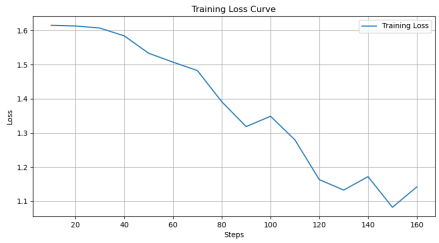

Figure 2: Training Loss Curve of Fine-tuning GuwenBERT for Task 2

5.3.3 Meta Llama-3 Chat (70B)

We used the API of the model through the Together AI platform. We set `max_tokens=700` such that all tokens in the prompts will be considered. We set `temperature=0.1` for more deterministic results so that the predicted poet better represents the model's best prediction.

5.3.4 Human Evaluation

To evaluate human performance on the poet classification task, four graduate students from the Department of East Asian Languages and Cultures (including the author) manually answered 60 questions of Task 2. Task 2 is chosen here because the poets in this task are selected such that the distinctiveness between their authorness is maximized, making it a good candidate to differentiate human understanding and machine understanding of the concept.

5.4 Results

Table 4 shows the results of the experiments.

As expected, the trivial baseline achieves the lowest performance, the GuwenBERT baselines achieve the highest performance, with the performance of the chosen LLM in between. It is unexpected that, however, the chosen LLM does not perform significantly better than the trivial baseline.

Task	Model	Accuracy	F1 Score	Precision	Recall
1	GuwenBERT (no fine-tuning)	0.63	0.55	0.57	0.63
1	GuwenBERT (fine-tuned)	**0.69**	**0.64**	**0.70**	**0.69**
1	Llama-3 Chat 70B (one-shot)	0.45	0.39	0.41	0.45
1	Llama-3 Chat 70B (five-shot)	0.47	0.39	0.44	0.47
1	Trivial	0.40	NA	NA	NA
2	GuwenBERT (no fine-tuning)	**0.58**	0.57	0.58	**0.58**
2	GuwenBERT (fine-tuned)	0.57	**0.58**	**0.61**	0.57
2	Llama-3 Chat 70B (one-shot)	0.31	0.28	0.35	0.31
2	Llama-3 Chat 70B (five-shot)	0.31	0.26	0.47	0.31
2	Trivial	0.27	NA	NA	NA
2	Human	0.53	NA	NA	NA

Table 4: Results of the experiments

6 Analysis

6.1 GuwenBERT Baselines

With respect to task 1, the fine-tuned version of GuwenBERT performs significantly better than the version without fine-tuning (in all metrics). However, for task 2, the performance of the fine-tuned version is not significantly better than that of the version without fine-tuning—for some metrics, the performance of the fine-tuned version is even slightly lower. One possible reason for this phenomenon might be that the poems for task 2 are already well-represented by the embedding of GuwenBERT, so fine-tuning would not significantly improve the accuracy of the embedding.

Note that the GuwenBERT baselines perform much better than the LLM and the trivial baseline in both tasks. For task 2, their accuracy is even higher than the human evaluation by graduate students in this field. This shows the promise of fine-tuning pre-trained models in the field of EADH.

6.2 LLM Performance

For both tasks, in comparison to the one-shot in-context training, the five-shot version does not significantly improve the performance of the LLM. Some possible reasons might be: First, the chosen poem might not represent the corresponding author's style very well, so it does not give the model crucial information to differentiate the poets. Second, the LLM might have already "learned" the information provided by the in-context tuning somewhere else, so the inclusion of these prompts does not add new information for the model. Third, the amount of in-context tuning data might be too small, and the model is unable to generalize well from the prompts. Fourth, the chosen LLM, Llama-3 Chat (70B), might not be pre-trained on classical Chinese corpora, so it does not have a deep understanding of the language of classical Chinese. It is possible that the model confuses classical Chinese and modern Chinese (it mistakenly applies its knowledge about modern Chinese to classical Chinese).

Note that even though the performance of the LLM is better than the trivial baseline, the improvement is not comparable to that from the LLM to the GuwenBERT baselines. The results suggest that the extensive application of LLM to the field of EADH might be premature at this moment.

It is possible that another LLM that includes classical Chinese corpora in its pre-training would perform much better than Llama-3 (or even fine-tuned GuwenBERT). However, training such an LLM would be expensive, and applications of LLMs in classical Chinese might not be as pragmatic as those in other modern languages. For example, LLMs trained on modern Chinese corpora might assist companies, governments, and organizations in the decision-making process. By contrast, LLMs trained on classical languages are more likely to be used in fields such as the preservation of cultural relics.

7 Conclusion

7.1 Main Findings

In this project, we investigate the potential of using LLMs in data augmentation for East Asian Digital Humanities. In particular, we choose the task of authorship attribution for classical Chinese poetry and test the performance of Llama-3 Chat 70B in comparison with fine-tuned GuwenBERT and a trivial baseline. The results show that the application of generative LLMs to constructing high-quality datasets for East Asian humanities might be premature. Instead, the use of fine-tuned transformers is more promising.

7.2 Limitations and Future Work

The main limitation of this project is the scope of experiments. In terms of the task chosen, authorship attribution is only one task among many other possible tasks in the field of EADH. It is possible that LLMs are better at handling other easier tasks and could be applied there. In terms of the model chosen, although Llama-3 Chat 70B is among the best models currently, it is still possible that other LLMs might perform better than the version used in this project. Also, only ten poets are chosen, and some of the poems might already be learned in the pre-training stage of the model.

For a more comprehensive study, more work could be done in these directions: First, we could evaluate more types of authorship attribution tasks, not only restricted to poetry. Since the vocabulary of poetry is smaller than, say, that of novels, LLMs might perform better in authorship attribution with respect to novels. Second, we could test more LLMs, and fine-tune some smaller LLMs to see whether they would achieve better results than the current one. Third, other techniques, such as transfer learning, could be applied to data augmentation for EADH. These avenues are also worth exploring.

8 Ethics Statement

One significant issue of authorship attribution based on machine learning is the confusion of authorship, causing misleading information in, say, educational contexts. Misattributing texts can undermine academic research and pose risks in education, where accuracy is crucial. Students might receive incorrect authorial attributions, distorting their understanding of literary traditions and authors' contributions. To mitigate this, government policies could restrict LLM use in education, ensuring only verified information is disseminated. Policies might involve mandatory verification of LLM outputs and expert-curated educational resources, preventing misinformation while allowing controlled LLM use.

Another issue of the task is the tendency of language models to favor canonical authors, marginalizing lesser-known and underrepresented figures. Extensive training data on well-known authors leads LLMs to prioritize these figures, skewing research and reinforcing disparities in the literary canon. To address this, human interference in training and application is essential. Curating balanced training datasets and involving human experts during training can prevent models from disproportionately favoring canonical authors. Human oversight can correct biases, promoting inclusive literary analysis.

References

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Ethan-yt. 2020. guwenbert. `https://github.com/Ethan-yt/guwenbert`.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

garychowcmu. 2019. Daizhigev20. `https://github.com/garychowcmu/daizhigev20`.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

jackeyGao. 2024. chinese-poetry. `https://github.com/chinese-poetry/chinese-poetry`.

Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddia, Nabil Arhab, et al. 2024. A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms. *arXiv preprint arXiv:2404.00303*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023a. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023b. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192.

M Novikova and P Novikov. 2024. Digital humanities: New research trajectories, new challenges, new interpretation. In *INTED2024 Proceedings*, pages 3155–3159. IATED.

Andrew Schonebaum. 2022. Approaches to teaching the plum in the golden vase (the golden lotus). Modern Language Association.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yuxuan Yao, Sichun Luo, Haohan Zhao, Guanzhi Deng, and Linqi Song. 2024. Can llm substitute human labeling? a case study of fine-grained chinese address entity recognition dataset for uav delivery. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1099–1102.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Jing Zhang, Hui Gao, Peng Zhang, Boda Feng, Wenmin Deng, and Yuexian Hou. 2024a. La-ucl: Llm-augmented unsupervised contrastive learning framework for few-shot text classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10198–10207.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024b. Llm–assisted data augmentation for chinese dialogue–level dependency parsing. *Computational Linguistics*, pages 1–24.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. Wyweb: A nlp evaluation benchmark for classical chinese. *arXiv preprint arXiv:2305.14150*.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*.