HUIXIANGDOU-CR: COREFERENCE RESOLUTION IN GROUP CHATS

Huanjun Kong[1,2]

[1]ncnn contributors
[2]Shanghai AI Laboratory

ABSTRACT

How to eliminate pronominal reference in group chats? In this work, we have preprocessed 58k authentic chat data and manually annotated 2.3k questions. The reliability of this annotation was confirmed by the scaling law. After this, we conducted fine-tuning on Qwen models, ranging from 0.5B to 32B parameters. The optimal version improved 29.07 in F1 score. This confirms the viability of fine-tuning Large Language Model (LLM) for downstream Natural Language Processing (NLP) tasks. Our contributions are: 1) Created Supervised Fine-Tuning (SFT) training data in alpaca format, along with a set of Low-Rank Adaptation (LoRA) weights, and 2) Developed a method for acquiring high-quality data leveraging scaling law principle. The script, raw data with alpaca format and experiments track are open-sourced on Github, HuggingFace and WandB. The privacy of the data involved has been authorized by users[1].

1 INTRODUCTION AND RELATED WORK

In HuixiangDou Kong et al. (2024), we have released an industrial-grade assistant to handle inquiries within group chats. We also build rejection pipeline with text2vec model and LLM scoring. However, ChatML format is limited to the roles of **bot**, **system**, and **user**, with prompts in HuixiangDou only consisted of individual user inputs, which does not take full advantage of the dialogue from other users. This limitation is particularly evident when dealing with conversation involving pronouns, as shown in conversation example 1.

> **Group Chat Example**
>
> **Alice:**
> Can mmpose be deployed on mobile phones?
> **Bob:**
> BTW, how to deploy it on TX2 ?

Figure 1: A conversational example containing pronouns, in which "it" needs to be disambiguated to "mmpose".

We have experimented NLP methods in HuixiangDou, but apart from the straightforward segmentation of a single question into multiple parts using part-of-speech (POS) tagging, other methods have proven to be of low cost-effectiveness. On one hand, the NLP implementation is complex; on the other hand, open-sourced models like HanLP He and Choi (2021) are unable to support both Chinese and English well. Additionally, the part-of-speech in Chinese is more complex than in English. For instance, the Chinese equivalent of the word "deploy" can function as a verb, adjective, or noun depending on the context. When faced with thousands of users, there are always badcases that break NLP rules, leading to extremely high maintenance costs for the system.

---

[1]SFT data and license comes from ncnn contributors group.

ReALM Moniz et al. (2024) distinguishes three coreference resolution task types in mobile scenario, claiming that with manual annotation and training, a fine-tuned model with 80 million parameters can rival the capabilities of GPT-4. Qin et al. (2022) attempts to address how Pre-trained Language Models (PLMs) can learn universal representations and effectively adapt to a wide range of NLP tasks. Therefore, we believe that fine-tuning LLM for downstream NLP tasks is feasible.

Li et al. (2024) introduced Instruction-Following Difficulty (IFD) and has demonstrated that by filtering out 5% of the data, then better accuracy can be achieved. It proves that the quality of SFT data is more important than quantity. This raises the question for us: **what is the objective criteria for data quality in this scenario?**

Both LoRA Hu et al. (2021) and LoRA+ Hayou et al. (2024) assume that the increments of weights are of low rank. According to inferences made on Kong (2023), LoRA only activates the pre-existing knowledge in PLM but can't learn new content. So, how can we prove that the data in this scenario meets the low-rank assumption?

2 METHODOLOGY

HuixiangDou has been active in many groups for a year. Given the diverse backgrounds of the ncnn contributors and the high complexity of the topics discussed, which range from hardware to software and possess a certain degree of universality, and the group's high level activity—87 messages per person per month on average—makes it ideal for experiments.

Consequently, we have preprocessed 58,000 raw chat lines from ncnn contributors group, following these two steps:

- **Concat** Concatenating consecutive messages from the same user based on timestamp. In post-hoc analysis, a person need to send 2 messages to articulate 1 question.
- **Filter** Utilizing KIMI and InternLM2 to assess whether a query is a question, since HuixiangDou is primarily concerned inquiry within the group.

After preprocess is completed, we have 2,302 inquiries.

Then we take scaling law as objective metric for manual annotation. **The larger the model's parameter, the better vanilla LLM performance on the annotated data will be.**
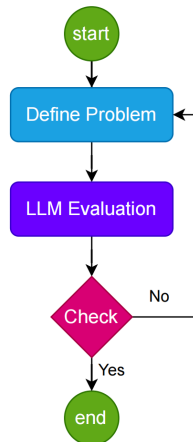

Figure 2: The overall structure of data annotation steps.

As shown in Figure 2, our annotation process involved the following steps:

1. **Problem Definition** Describe the task, manually annotate and define LLM prompt.
2. **LLM Evaluation** Take 7 Qwen LLMs (from 0.5B to 32B) as baseline and test them on the annotated data.

3. **Check** If precision does not increase with model size expanding, then revise the prompt and re-annotate, goto step 1.

After several iterations, the precision of the baseline models is shown in Table 1.

Model	Precision(%)	F1 score(%)
Qwen1.5-0.5B-Chat	**55.68**	52.43
Qwen1.5-1.8B-Chat	**56.15**	50.02
Qwen1.5-4B-Chat	**56.92**	62.75
Qwen1.5-7B-Chat	**57.11**	68.17
Qwen1.5-14B-Chat	**60.77**	64.59
Qwen1.5-32B-Chat	**68.29**	60.15
Qwen1.5-MoE-2.7B-Chat	61.79	32.86

Table 1: Qwen LLMs precision is tested on annotated data. Based on the objectivity of the scaling law, the blue part confirms the reliability of manual annotation. With the change in model architecture, Mixture of Experts (MoE) result is only for reference.

We fine-tuned 7 models on annotated data. Ultimately, Qwen1.5-MoE-2.7B-Chat F1 score significantly improved from 32.07% to 61.93%. Moreover, Qwen1.5-14B-Chat excelled in recall increase, from 68.91% to 92.11%.

3 DATASET

Our preprocessed group chat data with JSON format is shown below, with paramater explaination:

```
1  {
2    "id": INTEGER,
3    "sender": "STRING",
4    "text": "SRING",
5    "timestamp": INTEGER,
6    "kimi_is_question": BOOL,
7    "cr_window": [{
8      "sender": "STRING",
9      "text": "STRING",
10     "timestamp": INTEGER
11   }],
12   "cr_need_gt": BOOL
13 }
14
15 # sender: message sender ID, it is random STRING, not username
16 # is_question: filter by InternLM2 scoring, does this text is
17 a question
18 # kimi_is_question: filter by KIMI scoring, the throttle is 5
19 # cr_window: historical messages before the text, up to a maximum of 12
20 # cr_need_gt: manual annotation, does this text need resolution or not
```

As alpaca format below needs **input** or **instruction**, we have to merge **text** and **cr_window** into single sentence.

```
1  {
2    "instruction": "STRING",
3    "input": "STRING",
4    "output": "STRING"
5  }
```

In this scenario, even if coreference resolution fails, the original input is retained, allowing LLM making ternary decision to disambiguate the input, thus alpaca **instruct** includes single-choice question formats, such as "A. Not needed B. Needed C. Don't know".

3

Implementation Details To prevent overfitting, we randomly adjusted the correct option position in alpaca **output**, with some data having the true value in option A and others in option B or C, and the corresponding implementations are available on Github.

4 SFT EXPERIMENTS

We fine-tuned the Qwen series models using LoRA in axolotl contributors (2023), all configurations can be found on Github.

At the initial state, we set epoch to 1, learning rate to 2e-4 and LoRA rank to 64. The accuracy is detailed in Table 2.

Model	F1 score(%)	improvement(%)
Qwen1.5-0.5B-Chat	60.60	+8.17
Qwen1.5-1.8B-Chat	62.90	+12.88
Qwen1.5-4B-Chat	51.25	-11.50
Qwen1.5-7B-Chat	69.22	+1.05
Qwen1.5-14B-Chat	77.09	+12.50
Qwen1.5-32B-Chat	85.58	+25.43
Qwen1.5-MoE-2.7B-Chat	61.93	+29.07

Table 2: Finetuned Qwen models F1 score, the blue part also exhibits a trend in line with the scaling law, which validates the reliability of the annotation.

We observed a F1 score drop in Qwen1.5-4B-Chat model and we believe that a rapid decrease in loss might affect performance, so we updated the learning rate from 2e-4 to 2e-5. This adjustment effectively mitigated the decline in F1 score, which improved from -11.50 to -3.44. The new loss curve shown in Figure 3.
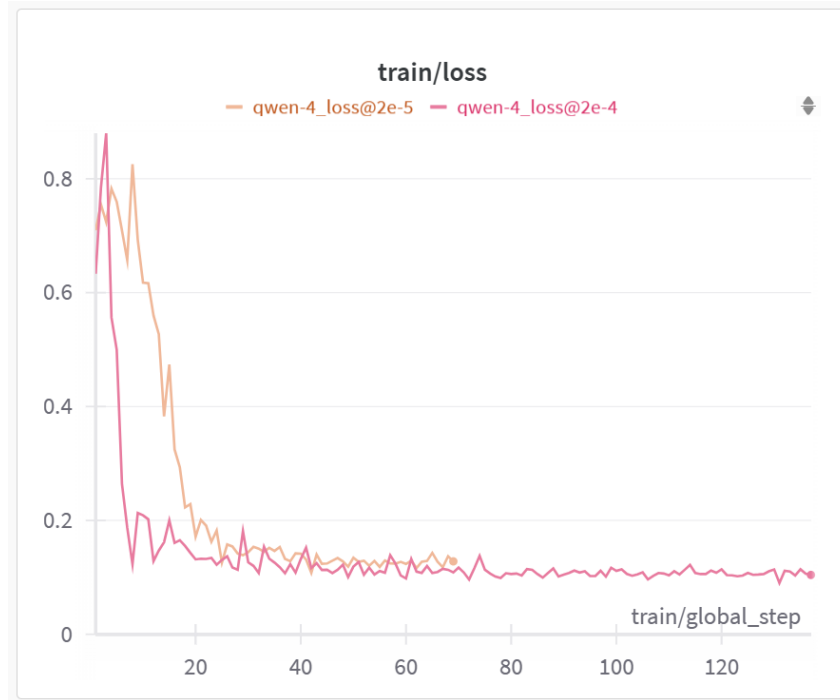

Figure 3: The curves representing different learning rates indicate that when Qwen1.5-4B-Chat learning rate set to 2e-5, the loss in F1 score is mitigated.

We also experimented with various rank size and LoRA+ for optimization, the results are presented in Table 3.

Method	F1 score(%)
vanilla	62.75
LoRA lr=2e-5	59.31
LoRA lr=2e-4	51.25
LoRA rank=16	53.27
LoRA rank=32	56.54
LoRA rank=64	51.25
LoRA rank=128	49.31
LoRA+ epoch=1	42.96
LoRA+ epoch=2	52.89
LoRA+ epoch=4	47.65

Table 3: Finetuned Qwen1.5-4B-Chat with different methods. Reducing learning rate can only mitigate the decline in accuracy, but it does not lead to an precision improvement.

5 CONCLUSION AND LIMITS

We take scaling law as guiding principle to define the problem and manually annotate the data. The model after SFT also conforms to the laws of scaling as shown in Figure 4, which proves the effectiveness of annotation.
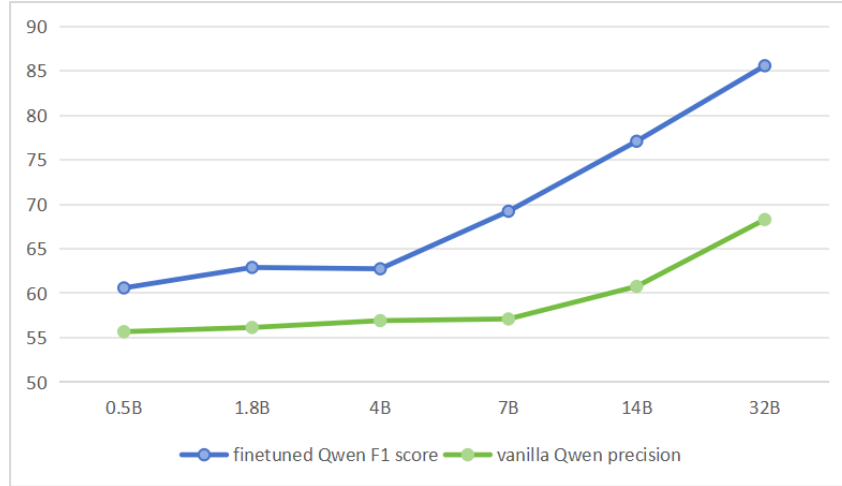

Figure 4: We compared the performance of the vanilla Qwen series with those fine-tuned models. The results showed that as the model parameters increased, the F1 scores and precision improved, which further confirmed the reliability of our dataset.

However, Qwen1.5-4B-Chat issue remains unresolved. We need to further prove that LoRA prerequisites are met. As a full-stack open-sourced project, HuixiangDou also needs to integrate the model and verify its overall effectiveness.

REFERENCES

Axolotl contributors. Axolotl is a tool designed to streamline the fine-tuning of various ai models, offering support for multiple configurations and architectures., 2023. URL `https://github.com/OpenAccess-AI-Collective/axolotl`.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models, 2024.

Han He and Jinho D. Choi. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.emnlp-main.451`.

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

Huanjun Kong. The unreliability of lora., Nov 2023. URL `https://zhuanlan.zhihu.com/p/662295224`.

Huanjun Kong, Songyang Zhang, Jiaying Li, Min Xiao, Jun Xu, and Kai Chen. Huixiangdou: Overcoming group chat scenarios with llm-based technical assistance, 2024.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning, 2024.

Joel Ruben Antony Moniz, Soundarya Krishnan, Melis Ozyildirim, Prathamesh Saraf, Halim Cagri Ates, Yuan Zhang, Hong Yu, and Nidhi Rajshree. Realm: Reference resolution as language modeling, 2024.

Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Jing Yi, Weize Chen, Zhiyuan Liu, Juanzi Li, Lei Hou, Peng Li, Maosong Sun, and Jie Zhou. Exploring universal intrinsic task subspace via prompt tuning, 2022.