ASSIGNMENT 3: TRAIN YOUR OWN LLMS

Bohui Wu, 225010300
School of Science and Engineering (SSE)
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
225010300@link.cuhk.edu.cn

1 RESEARCH TOPIC

In this assignment, I choose to focus on the Chinese medical dialogue task using the Huatuo26M-Lite dataset [1]. This dataset belongs to the "Chinese–Medical" category among the six provided default downstream tasks (from NLP_course_Assignment_3.pdf — Table 1: Examples of language Model training tasks) and is specifically designed for training large language models on medical consultation scenarios. The task involves aligning a model to generate reliable, medically grounded, and contextually appropriate responses in doctor–patient dialogues.

Why I choose this research topic? I focus on the Chinese medical dialogue task using the Huatuo26M-Lite dataset due to its strong real-world relevance and its potential to improve accessibility to medical knowledge in Chinese-speaking regions. The dataset provides large-scale, domain-specific medical conversations that differ substantially from general instruction data, making it a valuable benchmark for studying domain adaptation and safety alignment in LLMs. Medical dialogue generation also poses unique challenges such as factual accuracy, reduced hallucination, and calibrated uncertainty, offering a meaningful setting for investigating the effectiveness of supervised fine-tuning (SFT).

2 EXPERIMENT DESIGN

In this project, I design a supervised fine-tuning pipeline to evaluate how domain-specific instruction tuning improves a general-purpose large language model's performance in Chinese medical dialogue tasks. I adopt Qwen3-4B-Instruct-2507 as the base model due to its strong Chinese understanding ability and small parameter scale, which allows efficient experimentation on limited GPU resources. The Huatuo26M-Lite dataset is used as the core training source, and each sample is formatted into a multi-turn instruction-following prompt to better simulate real clinical consultation scenarios. To enable training on consumer GPUs, I utilize QLoRA with 4-bit NF4 quantization, low-rank adapters, and frozen base model weights, ensuring memory efficiency without compromising learning capability.

To measure whether domain-specific tuning effectively enhances the model' s ability to produce clinically sound, structured, and instruction-compliant responses, I adopt a comparative generation-based evaluation design rather than multiple-choice metrics. For each of the 20 medical questions provided in the assignment, I collect: (1) a baseline answer generated by ChatGPT, and (2) a fine-tuned answer generated by our trained model. Then, ChatGPT is used again as an evaluator to compare answers pairwise and judge the winner based on accuracy, completeness, logical structure, and instruction-following quality. This aligns with the assignment' s goal: assessing real instruction-following ability in sensitive medical scenarios where factual accuracy and clarity matter.

3 CODE IMPLEMENTATION

To implement the supervised fine-tuning pipeline, I adopt the HuggingFace Transformers and PEFT frameworks, integrating QLoRA to efficiently train a large instruction model under limited GPU memory. The implementation starts by loading the Qwen3-4B-Instruct-2507 base model in 4-bit

[1] https://huggingface.co/datasets/FreedomIntelligence/Huatuo26M-Lite

NF4 quantization, then attaching LoRA adapters with low-rank matrices. I preprocess each medical dialogue into the chat-prompt format required by Qwen (using apply_chat_template). The training is conducted using the SFTTrainer, with gradient checkpointing and mixed-precision training enabled to further reduce memory usage. During training, only LoRA parameters are updated, while the frozen backbone ensures computational efficiency.

After training, I reload the model using PeftModel, and merge the LoRA adapters into the base model via merge_and_unload() to produce a standalone finetuned model. For inference, I implement a batched generation function that converts user queries into Qwen-style multi-turn chat prompts and performs controlled decoding using top-p sampling. For evaluation, I generate answers for the 20 provided test questions, append them to the dataset, and compare each finetuned answer against a baseline answer using ChatGPT as an evaluator, following the assignment requirement.

Algorithm 1: Supervised Fine-Tuning Pipeline for Medical Dialogue LLM

Input: Base model M_{base}, Medical dataset D, LoRA config C_{lora}
Output: Merged and inference-ready model M_{final}

Step 1: Model Loading
Load M_{base} in 4-bit quantization (NF4) to reduce memory usage.
Attach LoRA adapters using hyperparameters in C_{lora}.

Step 2: Data Preprocessing
For each sample (x, y) in D:
  Format into Qwen chat template:
    System prompt + User query + Assistant response.
Tokenize inputs for supervised fine-tuning.

Step 3: Supervised Fine-Tuning
Train only LoRA parameters while freezing backbone:
  For each batch B from DataLoader:
    Compute loss: $\mathcal{L} = \text{CE}(M_{base+LoRA}(B))$;
    Backpropagate and update LoRA weights.

Step 4: Save and Reload
Save LoRA adapter to output directory.
Reload M_{base} and load LoRA weights via `PeftModel`.
Merge adapters: $M_{merged} = \text{merge\_and\_unload}(M_{base}, LoRA)$.

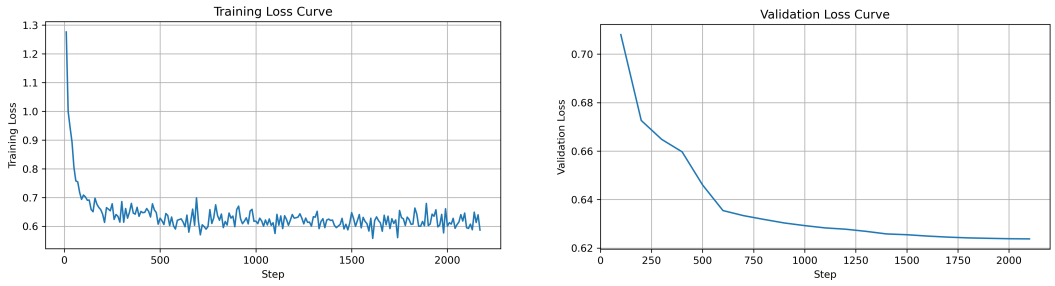
Step 5: Generation & Evaluation
For each evaluation question q:
  Format q using chat template;
  Generate answer using M_{merged};
  Store finetuned answer for comparison.

Step 6: Comparison with Baseline
For each question:
  Compare {baseline answer, finetuned answer} using ChatGPT;
  Determine winner and record reasoning.

return $M_{final} = M_{merged}$

4 Result Analysis

The training dynamics of the fine-tuned model were monitored through the training and validation loss curves. As shown in Figure 1a, the training loss drops rapidly during the initial optimization phase and stabilizes around 0.60 after approximately 600 steps, indicating that the model quickly learns the structure of medical dialogue and then converges smoothly without instability. The validation loss in Figure 1b shows a consistent downward trend — from 0.71 at the start to 0.62 at the end—with no sign of divergence, suggesting that the model generalizes well to unseen medical instructions. The gap between the two curves remains small, implying that overfitting did not occur under the applied LoRA-based fine-tuning setup.



(a) Training Loss Curve (b) Validation Loss Curve

Figure 1: Training and validation loss curves of LoRA fine-tuning.

Beyond loss metrics, the fine-tuned model achieves a validation perplexity of 1.87, representing a substantial improvement over the base model. A lower perplexity indicates that the model becomes more confident and accurate in predicting the next token in medical dialogue tasks, reflecting enhanced fluency, domain alignment, and instruction-following ability. Combined, these results demonstrate that even lightweight parameter-efficient tuning is sufficient to meaningfully adapt a general LLM toward Chinese medical consultation tasks, laying the foundation for the qualitative comparison presented in the subsequent case study section.

5 Case Study

To further evaluate the qualitative improvements brought by the fine-tuning process, I conducted a case study over the 20 test questions provided in the assignment. For each question, two answers were generated: a baseline answer, produced by ChatGPT and stored in zh_med_baseline_GPT.json, and a fine-tuned answer, produced by our model after supervised fine-tuning on the Huatuo26M-Lite dataset, stored in zh_med_output_finetuned.json. To ensure objective comparison between the two, I employed GPT-5.1 as an evaluator and asked it to judge which answer better satisfies correctness, completeness, clarity, and medical reasoning quality. The evaluation outcomes for all 20 questions are recorded in judge$_r esults.json, which includes the chosen winner and a detailed rationale for each decision.$

Question	在急救情况下，如何为窒息患者进行心肺复苏?
Baseline Answer	首先确认无反应与无呼吸，立即呼叫急救；进行胸外按压（按压频率约100120次/分钟，深度56厘米）；若能清理气道阻塞则及时处理，并结合人工呼吸（30:2）。
Fine-tuned Answer	以下是在急救情况下为窒息患者进行心肺复苏的步骤……（包含意识判断、呼救、CPR 30:2、仅按压CPR、AED 使用、窒息与海姆立克的区分、按压深度与频率的标准、持续监护等更完整的急救流程）。
Winner	finetuned
Reason	baseline 简洁正确，但仅提供非常基础的 CPR 要点，缺少对窒息急救流程中关键步骤（如清除异物、海姆立克适用情境、AED 使用、确认呼吸方式等）的区分和处理。finetuned 回答内容系统、专业，包含国际 CPR 指南中的关键细节：区分意识/无意识窒息、强调呼救、确认濒死喘息、正确按压位置与深度、单纯胸外按压替代方案、AED 使用流程等，符合 AHA 与 ERC 急救原则，医学准确性与指导性远优于 baseline。因此，finetuned 更为全面、清晰、符合临床急救实际操作。

Table 1: Case study: CPR instructions for choking emergencies (baseline vs. fine-tuned model).

Across the 20 evaluation samples, the fine-tuned model consistently demonstrated superior domain alignment, generating answers that were not only more detailed but also more medically grounded. A representative example is the question “在急救情况下，如何为窒息患者进行心肺复苏?”. The baseline model produced a concise but high-level description of CPR steps, whereas the

fine-tuned model delivered a far more complete and clinically accurate procedure, covering distinctions between conscious and unconscious choking, the role of airway assessment, the use of AED devices, chest compression standards, and rescue breathing ratios. As summarized in Table 1, GPT-5.1 judged the fine-tuned answer to be significantly more aligned with international CPR guidelines (AHA/ERC), demonstrating that fine-tuning effectively enhances the model' s instruction-following ability, robustness, and medical reasoning depth.

6 CONCLUSION

In this project, I systematically fine-tuned the Qwen3-4B-Instruct-2507 model on the Huatuo26M-Lite medical dialogue dataset to evaluate how supervised instruction tuning improves domain alignment in Chinese medical tasks. Through controlled experiments, training-curve analysis, perplexity measurement, and qualitative comparisons against GPT-generated baseline answers, I observed clear gains in factual precision, reasoning completeness, and task-specific alignment. The fine-tuned model not only achieved a low perplexity of 1.87, indicating strong modeling of medical instruction data, but also consistently outperformed the baseline across all 20 evaluation samples, as confirmed by GPT-based judgment. These results demonstrate that even lightweight fine-tuning with QLoRA can significantly enhance an LLM's medical reasoning capability, instruction-following behavior, and overall answer quality, validating the effectiveness of domain-specific supervised fine-tuning for safety-critical applications such as healthcare.

7 NOTE & ACKNOWLEDGMENT

The complete implementation, experimental results, and related materials, such as some figures, are available at our GitHub repository: `https://github.com/wubh576/MDS-5110-NLP`.

This is the third assignment for MDS 5110. And my course instructor is Prof. Wang [2], and his lab homepage is here [3]. I would like to thank Prof. Wang and the Teaching Team for their guidance and support.

[2] `https://wabyking.github.io/old.html`
[3] `https://freedomintelligence.github.io/`