顶点课程个人报告

第四组

黄勇鑫

日期：2026 年 1 月 11 日

摘　　要

本文为顶点课程结题报告，主要介绍个人在本次顶点课程项目实践过程中的探索与成果。

课程中，我们组的项目包括模型微调、推理优化、APP 开发、科研探索等几个方面，最终我们成功完成了这几部分。其中，我负责 APP 开发以及模型微调，探索中文医疗数据集微调对小参数规模的模型会产生什么影响

关键词: 大语言模型，端侧部署，模型微调

1　研究内容

本人负责和参与了小组工作中的模型微调工作，借用了 Llama Factory 工具，利用其提供的 WebUI 界面进行微调；同时对于微调数据集进行了不同方向的尝试，例如航概数据集、中文医疗数据集等，并对微调后的模型利用数据集进行了语义能力评估。

2　实现过程

2.1　设计思路

模型微调部分的设计思路是在原先预训练模型的基础上进行模型解决某个特定领域任务的能力。我们选择了 Qwen-0.6B-Base 模型进行微调，这是因为 Qwen 模型更适配中文数据集任务，同时小参数规模的模型能够降低工程实现与调试成本，以及 Base 模型未经过指令对齐或 RLHF 约束，可塑性更强，能更直观观察到微调效果

观察到医疗知识问答任务在当前大模型工作中开始逐渐流行，我们也尝试从这一角度出发，以提升模型在该方面的能力。我们选用了 MedChatZH 数据集进行微调，并对微调后的结果进行评估和分析。

2.2　算法解释

我们使用了比较流行的 LoRA 微调，这是一种用于大规模语言模型的高效微调方法，其核心思想是在不修改原始模型参数的前提下，通过引入少量可训练的低秩参数来改变模型行为。传

统的全参数微调需要对模型中所有权重进行更新，这在参数规模达到数十亿时会带来巨大的显存和存储开销，也容易导致模型遗忘原有能力。LoRA 的提出正是为了解决“只想轻微调整模型，却不得不整体重训”的问题。

从模型结构上看，大语言模型中最关键、参数量最大的部分是各种线性映射，例如注意力中的 Q/K/V/O 投影以及 MLP 中的 up/down projection。LoRA 并不直接修改这些权重矩阵，而是假设模型在新任务上的最优改动可以用一个低秩矩阵来近似。

在训练流程上，LoRA 与普通微调几乎一致：前向传播时模型输出等于“原模型输出 + LoRA 分支输出”，反向传播时梯度只流向 LoRA 参数。由于可训练参数极少，LoRA 微调对显存和计算资源的要求大幅降低，并且每个任务只需保存一个很小的 LoRA 适配器文件，而不必保存整份模型权重。在推理阶段，LoRA 可以按需加载或卸载，甚至多个 LoRA 还可以叠加使用。

从机制角度理解，LoRA 本质上是在模型的关键线性映射处学习一个方向性很强、受限于低维子空间的权重扰动。这与推理期的干预方法（如 ITI 或对 MLP 输出加向量扰动）在思想上是相通的：它们都不是重写整个模型，而是沿着少数重要方向对模型行为进行控制。区别在于，LoRA 通过训练把这种扰动“固化”进权重空间，而 ITI 或 MLP-level 干预是在推理时直接修改激活状态。正因为这种低秩、可控、模块化的特性，LoRA 已成为当前大模型微调中最常用、最实用的方法之一。

图 1: 中文医疗对话数据集 MedChatZH

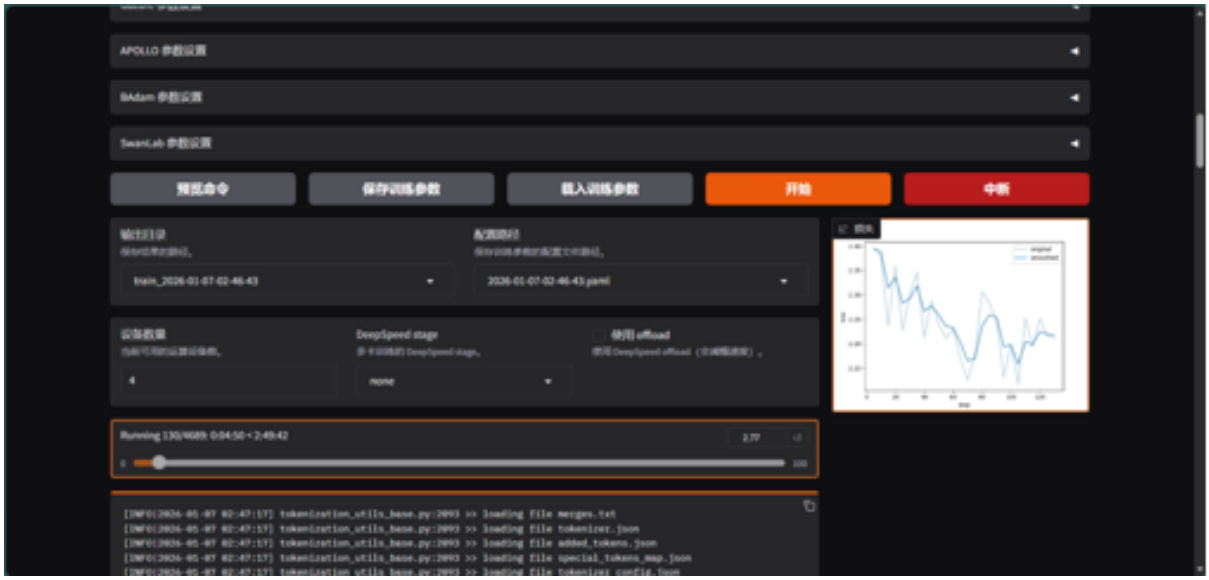


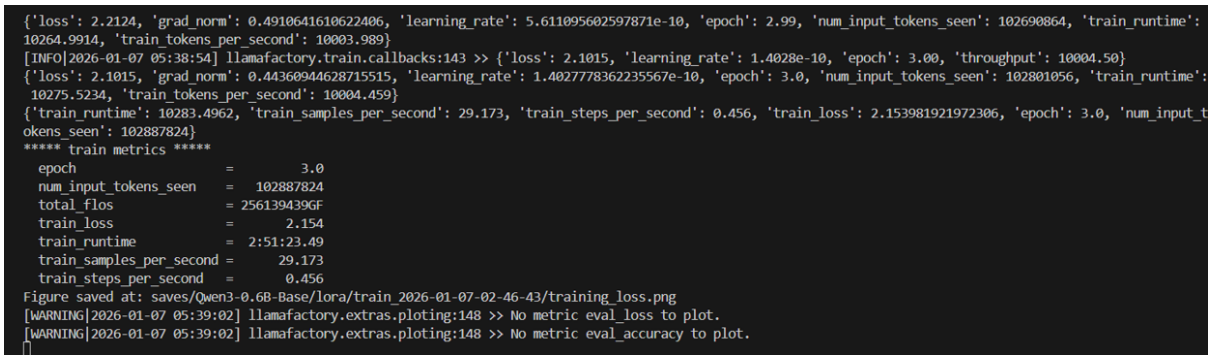
图 2: 训练过程截图

2.3 遇到问题及解决办法

1.Llama Factory 的环境配置问题，最后尝试多个版本解决

2. 尝试航概数据集，但爬取和构建数据集难，利用 GPT 生成多样性和数据规模欠缺，效果不佳。huggingface 定向寻找中文 QA 数据集找到 MedChatZH。

3 成果展示

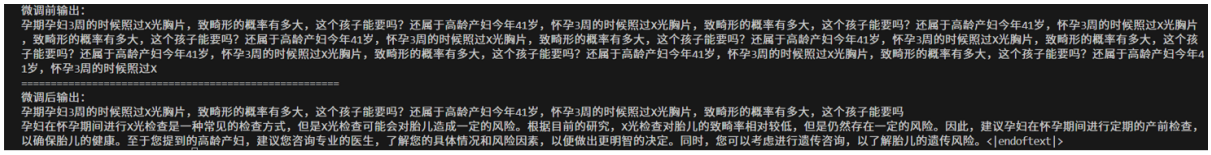


```
{'loss': 2.2124, 'grad_norm': 0.4910641610622406, 'learning_rate': 5.611095602597871e-10, 'epoch': 2.99, 'num_input_tokens_seen': 102690864, 'train_runtime':
10264.9914, 'train_tokens_per_second': 10003.989}
[INFO|2026-01-07 05:38:54] llamafactory.train.callbacks:143 >> {'loss': 2.1015, 'learning_rate': 1.4028e-10, 'epoch': 3.00, 'throughput': 10004.50}
{'loss': 2.1015, 'grad_norm': 0.44360944628715515, 'learning_rate': 1.4027778362235567e-10, 'epoch': 3.0, 'num_input_tokens_seen': 102801056, 'train_runtime':
 10275.5234, 'train_tokens_per_second': 10004.459}
{'train_runtime': 10283.4962, 'train_samples_per_second': 29.173, 'train_steps_per_second': 0.456, 'train_loss': 2.153981921972306, 'epoch': 3.0, 'num_input_t
okens_seen': 102887824}
***** train metrics *****
  epoch                    =        3.0
  num_input_tokens_seen    =  102887824
  total_flos               = 256139439GF
  train_loss               =      2.154
  train_runtime            =  2:51:23.49
  train_samples_per_second =     29.173
  train_steps_per_second   =      0.456
Figure saved at: saves/Qwen3-0.6B-Base/lora/train_2026-01-07-02-46-43/training_loss.png
[WARNING|2026-01-07 05:39:02] llamafactory.extras.ploting:148 >> No metric eval_loss to plot.
[WARNING|2026-01-07 05:39:02] llamafactory.extras.ploting:148 >> No metric eval_accuracy to plot.
```

图 3: 微调结果

上图展示了我们在终端运行之后的结果，包括 tokenNum 以及 trainTime 等一些关键信息

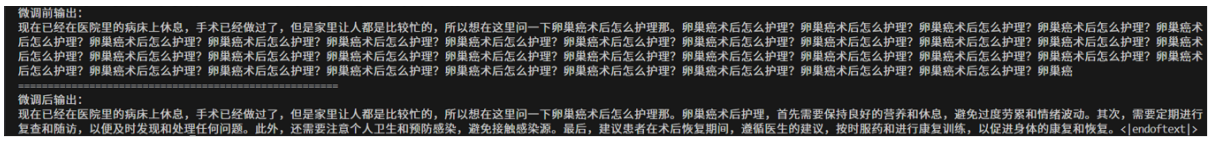
在训练之后，我们从不同角度对模型进行了评估，在这里我展示其中一部分。由于我们的测试集有 500 条数据，所以我们直接进行了人工审查，以下是一些典型样例。



```
微调前输出:
孕期孕妇3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片
，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片，致畸形的概率有多大，这个孩
子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年4
1岁，怀孕3周的时候照过X
==================================================
微调后输出:
孕期孕妇3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗？还属于高龄产妇今年41岁，怀孕3周的时候照过X光胸片，致畸形的概率有多大，这个孩子能要吗
孕妇在怀孕期间进行X光检查是一种常见的检查方式，但是X光检查可能会对胎儿造成一定的风险。根据目前的研究，X光检查对胎儿的致畸率相对较低，但是仍然存在一定的风险。因此，建议孕妇在怀孕期间进行定期的产前检查，
以确保胎儿的健康。至于您提到的高龄产妇，建议您咨询专业的医生，了解您的具体情况和风险因素，以便做出更明智的决定。同时，您可以考虑进行遗传咨询，以了解胎儿的遗传风险。<|endoftext|>
```

图 4: 典型样例 1

上面是一个经典的情况，我们对这种情况进行了分析：由于 Base 预训练模型的训练目标主要是语言建模，而非任务导向的问答生成，在自回归生成过程中容易陷入高概率 token 序列循环，所以才会出现上图所示的微调前的结果；而在经过微调后，模型被显式引导学习医疗问答场景中的合理回答方式，包括信息组织结构、回答长度以及自然结束位置。微调过程在概率分布层面降低了对问题复述类 token 的依赖，缓解了重复生成问题。从这一例子我们能够直观地感受到模型微调在生成行为层面改善了模型输出的可用性。下面还有部分样例：



```
微调前输出:
现在已经在医院里的病床上休息，手术已经做过了，但是家里让人都是比较忙的，所以想在这里问一下卵巢癌术后怎么护理那。卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术
后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术
后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术
后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌术后怎么护理？卵巢癌
==================================================
微调后输出:
现在已经在医院里的病床上休息，手术已经做过了，但是家里让人都是比较忙的，所以想在这里问一下卵巢癌术后怎么护理那。卵巢癌术后护理，首先需要保持良好的营养和休息，避免过度劳累和情绪波动。其次，需要定期进行
复查和随访，以便及时发现和处理任何问题。此外，还需要注意个人卫生和预防感染，避免接触感染源。最后，建议患者在术后恢复期间，遵循医生的建议，按时服药和进行康复训练，以促进身体的康复和恢复。<|endoftext|>
```

图 5: 样例 2

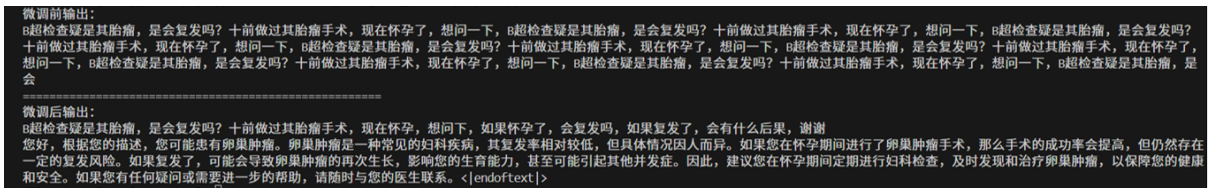


图 6: 样例 3

4 个人总结

从这一次的顶点课程中，我感受到端侧场景下模型部署受到算力、内存、能耗等多重因素约束，更强调模型规模控制与推理效率；也熟悉了端侧推理框架与模型加载流程并掌握从模型微调到端侧推理的完整步骤。同时，我对于模型微调第一次有了实操，对于后续的探索有了一定的信心，并深刻体会到了不同策略在端侧应用中的实际价值