CCL25-Eval任务8系统报告：基于规则奖励与自主思考强化学习的中文电子病历ICD诊断编码探索

邹游，张蕾，梁晓东，莫坤东，郭子滔，危枫，王晨子
中国电信股份有限公司重庆分公司，重庆
{zouy12, zhangl157, liangxd2, mokd, guozt, weifeng2.cq, wangchenzi.cq}
@chinatelecom.cn

摘要

世界卫生组织国际疾病分类ICD诊断编码的自动生成是医疗信息化的核心挑战，面临主诊断单标签分类准确性不足、其他诊断多标签预测不完整以及长尾分布等技术瓶颈。本文系统研究探索了大语言模型在中文电子病历ICD诊断编码任务中的微调范式创新，针对生成式微调、判别式微调，以及强化学习分别提出了不同的微调训练策略。其中，创新性地设计针对医疗特性的基于规则奖励的强化学习框架（RBRs-RL），通过动态难度校准、Token级梯度优化和超长奖励塑造策略改进了GRPO算法的效率和性能，同时结合提出的策略轮动数据增强迭代训练（SRADIT）策略，实现了强化微调性能上限的提升。此外，本文还系统比较了生成式与判别式微调在中文诊断ICD编码任务中的性能边界，同时构建了端到端的临床决策优化框架，为奖励微调提供有效路径。并且针对推理阶段，本文设计了一种温度调节集成共识预测方法（TCECP），提升了推理的稳定性和可靠性。最后基于Qwen2.5-7B模型的微调实验结果表明，通过本文提出的优化后的RBR-R1式强化微调方法，在CCL25-Eval任务8的A榜和B榜分别取得80.98和82.33的优异成绩，其效果显著超越传统SFT的性能上限。综上所述，本文的探索与发现为医疗诊断编码系统的实际应用提供了重要的技术参考。

关键词： 大语言模型；ICD编码；强化学习；GRPO

System Report for CCL25-Eval Task 8: Exploration of ICD Diagnosis Coding in Chinese EMRs Based on Rule-Based Rewards and Autonomous Thinking Reinforcement Learning

You Zou, Lei Zhang, Xiaodong Liang, Kundong Mo, Zitao Guo, Feng Wei, Chenzi Wang
China Telecom Corporation Limited Chongqing Branch, Chongqing
{zouy12, zhangl157, liangxd2, mokd, guozt, weifeng2.cq, wangchenzi.cq}
@chinatelecom.cn

Abstract

The automatic generation of World Health Organization International Classification of Diseases (ICD) diagnostic codes is a core challenge in medical informatization, facing technical bottlenecks such as insufficient accuracy in single-label classification of principal diagnoses, incomplete multi-label prediction of other diagnoses, and long-tail distribution. This paper systematically explores innovations in fine-tuning paradigms of Large Language Models (LLMs) for ICD diagnostic coding tasks in Chinese Electronic Medical Records (EMRs), proposing distinct fine-tuning strategies for generative

第二十四届中国计算语言学大会论文集，第322页-第330页，济南，中国，2025年8月11日至14日。

fine-tuning, discriminative fine-tuning, and Reinforcement Learning (RL). Notably, a medical features oriented Rule-Based Rewards Reinforcement Learning (RBRs-RL) framework is innovatively designed, which enhances the algorithm's efficiency via dynamic difficulty calibration, token-level gradient optimization, and overlong reward shaping strategies. Additionally, the proposed Strategy Rotation Data Augmentation Iterative Training (SRADIT) strategy enhances the upper performance limit of RL fine-tuning. This study also systematically compares the performance boundaries between generative and discriminative fine-tuning methods in Chinese diagnostic ICD coding tasks and establishes an end-to-end clinical decision optimization framework to provide an effective pathway for reward fine-tuning. In addition, this paper designs a Temperature-Calibrated Ensemble Consensus Prediction (TCECP) method, which enhances the stability and reliability of the inference phase. Finally, fine-tuning experiments based on the Qwen2.5-7B model demonstrate that the optimized RBR-R1 RL fine-tuning method achieves excellent results of 80.98 and 82.33 on the A and B leaderboards of CCL25-Eval Task 8, significantly surpassing the performance ceiling of traditional Supervised Fine-Tuning (SFT). Collectively, the explorations herein provide critical technical references for the practical application of medical diagnostic coding systems.

Keywords: Large Language Models , International Classification of Diseases (ICD) Codes , Reinforcement learning , Group Relative Policy Optimization (GRPO)

1 引言

随着医疗信息化进程的加速，电子病历系统中诊断编码自动化成为了医疗信息化的核心挑战，其准确性直接影响临床决策、医疗结算和流行病研究。世界卫生组织制定了国际疾病分类标准（International Classification of Diseases，ICD）编码体系作为全球通用的医疗数据标准，但在中文电子病历自动编码场景下面临三大技术瓶颈：1）主诊断的单标签分类需兼顾临床准确性与编码规范性；2）其他诊断的多标签预测存在长尾分布与标签不完整问题；3）传统基于规则的方法泛化能力不足，而深度学习方法难以平衡特征共享与任务特异性的需求 (Teng et al., 2023)。

大语言模型（Large Language Model，LLM）的出现为病历诊断ICD编码任务提供了新范式，然而现有研究存在三个关键局限：首先，主流工作过度依赖判别式微调，忽视了生成式方法在复杂临床推理中的优势；其次，多任务协同训练中主诊断与其他诊断的交互机制尚未系统探索；第三，强化学习在医疗编码任务中的奖励函数设计与训练稳定性问题亟待解决。这些局限导致现有方法在CCL25-Eval任务8等权威评测中表现受限。

针对上述挑战，我们系统探索了LLM在中文电子病历ICD自动编码任务中的微调范式创新。根据CCL25-Eval任务8的评测要求，我们提出了三种技术路线：1）在生成式微调中引入多任务协同训练(Multi-Task Learning，MTL)与思维链蒸馏(Chain-of-Thought Distillation，CoTD) (Wei et al., 2022)以及K-Fold多模型融合机制；2）在判别式微调中融合随机权重平均(Stochastic Weight Averaging，SWA) (Pavel et al. , 2019)与对抗训练(Adversarial Training，AT)来提升模型的鲁棒性；3）设计基于规则奖励的强化学习(Rule-Based Rewards Reinforcement Learning，RBRs-RL)框架实现临床偏好对齐与模型自主思考（Think）能力优化，并设计多种策略提升强化学习训练的效率稳定性以及上限。针对本次任务，我们选择了Qwen2.5-7B作为基座模型，并通过本技术报告提出的RBRs-RL微调方法实现了超越SFT方法的效果，在A榜以80.98分排名第二；在B榜以82.33分的成绩，取得了更优的效果。

本技术报告的贡献可归纳为：

1）系统比较了生成式与判别式微调在中文电子病历ICD诊断编码任务中的性能边界，提出的K-Fold投票集成机制显著提升了诊断多标签预测的稳定性和完整性；

2）设计医疗特异性的强化学习奖励函数体系，通过动态难度校准、Token级梯度优化和超长奖励塑造（Overlong Reward Shaping）策略，提升群体相对策略优化算法（Group Relative Policy Optimization，GRPO） (Shao et al. , 2024)的训练效率和性能；

第二十四届中国计算语言学大会论文集，第322页-第330页，济南，中国，2025年8月11日至14日。

3）设计策略轮动数据增强迭代训练（Strategy Rotation Augmented Data Iteration Training，SRADIT）策略，使用两阶段的数据增强机制，通过“识别-生成-验证-增强”的闭环流程，进行迭代式训练，提升强化微调的性能上限；

4）构建端到端的临床决策优化框架，为基于结果奖励模型（Outcome Reward Model，ORM）的微调提供更好的优化路径，推理阶段设计一种温度调节集成共识预测方法（Temperature-Calibrated Ensemble Consensus Prediction，TCECP），提升了推理阶段的稳定性和可靠性。让LLM在中文电子病历ICD诊断编码任务的性能上限突破了SFT微调方法。

上述技术探索为医疗诊断编码系统的实际应用提供了重要的实践参考。

2 方法

本次任务给定一段由临床信息构成的文本作为输入，需要模型输出对应的主诊断编码和其他诊断编码。该任务可以转化为两个子任务：

（1）通过病历信息生成一个主诊断，可以视为一个多分类任务；

（2）通过病历信息生成多个其他诊断，可以视为一个多标签分类任务。

针对这两个子任务，首先，通过引入MTL、CoTD和K-Fold集成学习等机制的方法对LLM进行生成式微调；其次，通过融合SWA与AT的判别式微调方法来提升LLM的鲁棒性；最后，我们设计了RBRs-RL框架来实现临床偏好对齐与模型自主Think。

2.1 微调样本设计

鉴于微调样本设计对模型优化效能有着显著影响，我们对微调数据集进行了系统化的样本设计工程（Sample Design Engineering，SDE）（Guo et al., 2024）。首先，我们对原始样本实施了诊断编码的语义转化工程，基于国家临床版2.0疾病诊断编码体系，将数据实体中的主要诊断与其他诊断标签，精准映射为规范化的疾病文本表征形式。在诊断特征的Prompt构建环节，针对主要诊断，运用专业医学词典，将五类典型疾病的临床表征摘要嵌入Prompt输入空间；对于其他诊断，则采用轻量化处理策略，仅将候选诊断列表集成至Prompt的特征向量空间。

针对R1式强化微调，主诊断生成任务和其他诊断生成任务的Prompt示例分别如图1和图2所示。

$Prompt$=“你是一个中文疾病诊断分类的专家，你会接收到一段包含患者就诊信息的文本和几个潜在的主诊断疾病，请输出文本内容的正确主诊断疾病，注意主诊断疾病仅有一个。

要求:请你先思考再给出答案。思考部分以< $think$ >开头，以< /$think$ >结尾，答案部分以< $answer$ >开头，以< /$answer$ >结尾。

主诊断疾病类别:

{main_sick_name}

高血压病3级（极高危）的定义：高血压病3级（极高危）是指收缩压≥ $180mmHg$和/或舒张压≥ $110mmHg$，且伴有多个心血管危险因素、靶器官损害或临床并发症的情况。极高危状态意味着患者在未来10年内发生心血管事件的风险极高。不稳定型心绞痛的定义：不稳定型心绞痛（UA）是急性冠脉综合征（ACS）的一种表现形式，属于冠心病的一种急性加重状态。其特点是心肌缺血症状在静息或轻微活动时突然发作，且症状较稳定型心绞痛更为严重、持续时间更长或发作频率增加。不稳定型心绞痛提示冠状动脉斑块破裂、血栓形成或血管痉挛，可能导致心肌梗死的风险显著增加。

微血管性心绞痛：微血管性心绞痛（MVA），又称心脏X综合征，是一种由冠状动脉微血管功能障碍引起的心绞痛。患者通常表现为典型的心绞痛症状，但冠状动脉造影显示主要冠状动脉无明显狭窄。MVA的病理机制涉及微血管内皮功能异常、血管平滑肌功能障碍及自主神经调节异常，导致心肌缺血。急性非ST段抬高型心肌梗死：急性非ST段抬高型心肌梗死（$NSTEMI$）是急性冠脉综合征（ACS）的一种，主要由冠状动脉部分阻塞引起的心肌缺血和坏死。与ST段抬高型心肌梗死（$STEMI$）不同，$NSTEMI$的心电图上不出现ST段抬高，但可能有ST段压低或T波倒置。慢性心功能不全急性加重：慢性心功能不全急性加重（$ADHF$）是指慢性心功能不全（CHF）患者因各种诱因导致心功能急剧恶化，出现明显的症状和体征，通常需要紧急医疗干预。$ADHF$是CHF患者常见的急性临床事件，可能导致住院甚至死亡。

患者就诊病案信息:

{info} ”

图 1: 主诊断Prompt格式示例

$Prompt$=“你是一个中文疾病诊断分类的专家，你会接收到一段包含患者就诊信息的文本和几个潜在的其他诊断疾病，请输出文本内容的正确其他诊断疾病。
要求:请你先思考再给出答案。思考部分以< $think$ >开头，以< /$think$ >结尾，答案部分以< $answer$ >开头，以< /$answer$ >结尾。
其他诊断疾病类别:
{other_sick_name}
患者就诊病案信息:
{info}”

图 2: 其他诊断Prompt格式示例

其次，我们开展了数据集的分层构建工程，遵循5-Fold交叉验证的统计学习范式，将原始训练集划分为五个互斥的子样本集，留存剩余数据构成独立测试集。针对每个子样本集进行模型微调训练，并对相应测试集实施预测效能评估，将预测偏差样本纳入误差样本库进行深度分析。

最后，针对误差样本库中的数据实体，我们启动了智能数据增强机制（Badcase Data Augmentation，BDA），使用LLM执行同标签病历重构任务，基于拒绝采样策略生成候选样本。首先，让LLM生成20个语义等效但表述相异的病历样本；随后，通过微调LLM进行标签一致性过滤；最终，由专家人工评分筛选最优增强样本，将其融入训练数据体系后，对LLM进行二次精细化微调，形成闭环优化的样本迭代机制。

2.2 生成式微调策略

为了探索生成式微调方法在中文电子病历ICD诊断编码任务中的潜力，我们提出了以下策略：

（1）多任务MTL：通过Prompt工程将病历信息编码为符合LLM输入规范的结构化序列，同步构建包含主诊断结果与其他诊断结果的复合标签空间。在此基础上实施多任务联合微调，赋能LLM实现主诊断与关联的其他诊断结果的协同输出能力构建，形成一体化的诊断决策生成框架。

（2）子任务SFT：采用任务解耦策略，将主诊断结果与其他诊断结果分别定义为Label1与Label2两个独立标签空间。通过Prompt技术对病历信息进行标准化编码后，针对两个子任务实施独立的SFT微调训练，构建专业化的子任务决策模型，实现诊断任务的精细化分工与模型能力的定向强化。

（3）子任务CoTD-SFT：使用Qwen2.5-7B模型，结合label生成其在处理病历信息时的完整CoT推理轨迹（Think Process）。针对主诊断与其他诊断两个子任务，分别实施CoT推理数据的定向生成，生成包含中间推理步骤的CoT数据（CoT Data）。通过将推理轨迹数据与原始任务数据深度融合，构建带有认知推理过程的监督微调语料，实现基于逻辑推导的诊断能力增强训练。

（4）子任务K-Fold交叉验证的SFT：采用5-Fold交叉验证框架构建鲁棒性训练体系，将病历信息通过Prompt编码后输入模型，分别以主诊断结果与其他诊断结果作为独立标签空间实施多轮次微调训练。在主诊断任务推理阶段，构建多数投票决策机制，通过5个子模型的预测结果集成提升主诊断结论的可靠性；针对其他诊断任务，为优化F1指标，形成多诊断结果集合，实现LLM输出的不确定性控制与结果稳定性优化。

2.3 判别式微调策略

为了比较生成式与判别式微调在中文电子病历ICD诊断编码任务中的性能边界，我们针对判别式微调构建了标准化的标签映射体系，将ICD编码系统映射为整数标签空间，为LLM微调训练提供统一的语义表示基础。在微调过程中，融合SWA技术与AWP (Yu et al., 2022)对抗训练。前者通过模型参数的动态平均机制提升输出稳定性，后者通过注入对抗扰动增强模型的鲁棒性。针对主诊断子任务，在LLM输出层构建专用分类器模块，实施端到端的诊断决策训练；对于其他诊断子任务，采用Softmax分类器结合交叉熵损失函数的经典架构 (Su et al. , 2022)，

通过优化类别概率分布提升多标签分类性能，构建兼具准确性与鲁棒性的判别式ICD诊断模型。

2.4 强化微调策略

针对两个子任务，我们设计了基于RBR的偏好学习强化微调，以及融入深度Think Process的RBR-R1式强化微调，来实现临床偏好对齐与模型自主Think能力优化。我们所采用的强化学习算法体系涵盖REINFORCE++(Hu et al. , 2025)、GRPO及其改进变体，构建了层次化的策略优化架构。

2.4.1 基于RBR的偏好学习强化微调机制

在基于RBR的偏好学习强化微调中，针对不同任务特性实施差异化算法配置：对主诊断子任务采用REINFORCE++算法，对其他诊断子任务则采用GRPO算法，构建标签级偏好强化学习范式。其中，奖励函数设计遵循精准度量原则：

主诊断子任务：通过精准比对模型输出与真实标签（ground truth），构建二元奖励机制。当输出结果完全匹配ground truth时，赋予1.0的完整奖励值；若输出与ground truth之间存在偏差，则奖励值归零，形成严格的正确性判别标准。

其他诊断子任务：引入F1-Score作为核心度量指标，将模型输出的综合性能量化为连续性奖励信号，实现对部分正确输出的梯度引导。

2.4.2 基于Think的RBR-R1式强化微调机制

在嵌入了深度Think Process的RBR-R1式强化微调中，我们采用GRPO算法构建策略优化体系，设计多维度奖励函数：

主诊断子任务：首先验证模型推理路径与答案是否符合预设的格式范式，若满足则赋予0.1的基础格式奖励；继而进行标签正确性校验，完全匹配ground truth时追加0.9的精准奖励，形成“格式+结果”的双重约束机制。

其他诊断子任务： 沿用主诊断子任务的格式奖励体系，同时将F1-Score与0.9的权重系数相乘，构建性能加权奖励函数，实现对输出质量的多维度量化评估。

$$J_{GRPO}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}[min(R_{qt}\hat{A_{i,t}}, clip(R_{qt}, 1-\epsilon, 1+\epsilon)\hat{A_{i,t}}) - \beta D_{KL}[\pi_\theta \parallel \pi_{ref}]] \quad (1)$$

$$R_{qt} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})} \quad (2)$$

GRPO算法的损失函数如公式（1）和（2）所示。其中，$clip(R_{qt}, 1-\epsilon, 1+\epsilon)$ 通过将策略比率限制在$1-\epsilon$和$1+\epsilon$之间，来确保更新不会过度偏离参考策略。针对GRPO算法在训练中呈现的收敛动态特性，我们提出经验数据筛选机制：在模型训练后期，当任务的批次正确率或F1-Score进入稳定区间时，对每次Rollout生成的经验数据进行难度校准。若批次数据出现标签全对或全错的极端情况，则判定为无效经验并舍弃，通过动态过滤简单/过难样本，保留中等难度经验子集，能够显著提升训练效率与模型性能上限。

在超长文本输出处理方面，我们采用基于连续函数的惩罚机制，基于超长奖励塑造（Overlong Reward Shaping）策略，构建平滑过渡的长度惩罚函数，如公式（3）所示。

$$R_{length}(y) = \begin{cases} 0 & if|y| \le L_{max} - L_{cache} \\ \frac{(L_{max}-L_{cache})-|y|}{L_{cache}} & if other \\ -1 & if|y| \ge L_{max} \end{cases} \quad (3)$$

其中，$L_{cache} = 6500$为缓冲区域，$L_{max} = 7000$为最大长度，y为输出长度。该机制通过函数衰减替代传统阶跃惩罚，能够有效规避奖励噪声问题。针对训练后期长Think序列的梯度贡献不足问题，我们将损失函数的计算粒度从样本级精细化为Token级，提高长Think序列对梯度的贡献。如公式（4）所示，通过对每个Token的贡献进行平等加权，显著增强复杂长序列对策略更新的有效引导，从而实现深度Think能力的定向优化。

$$\hat{J}_{GRPO}(\theta) = E_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}[min(R_{qt}\hat{A}_{i,t}, clip(R_{qt}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}) - \beta D_{KL}[\pi_\theta \parallel \pi_{ref}]] \tag{4}$$

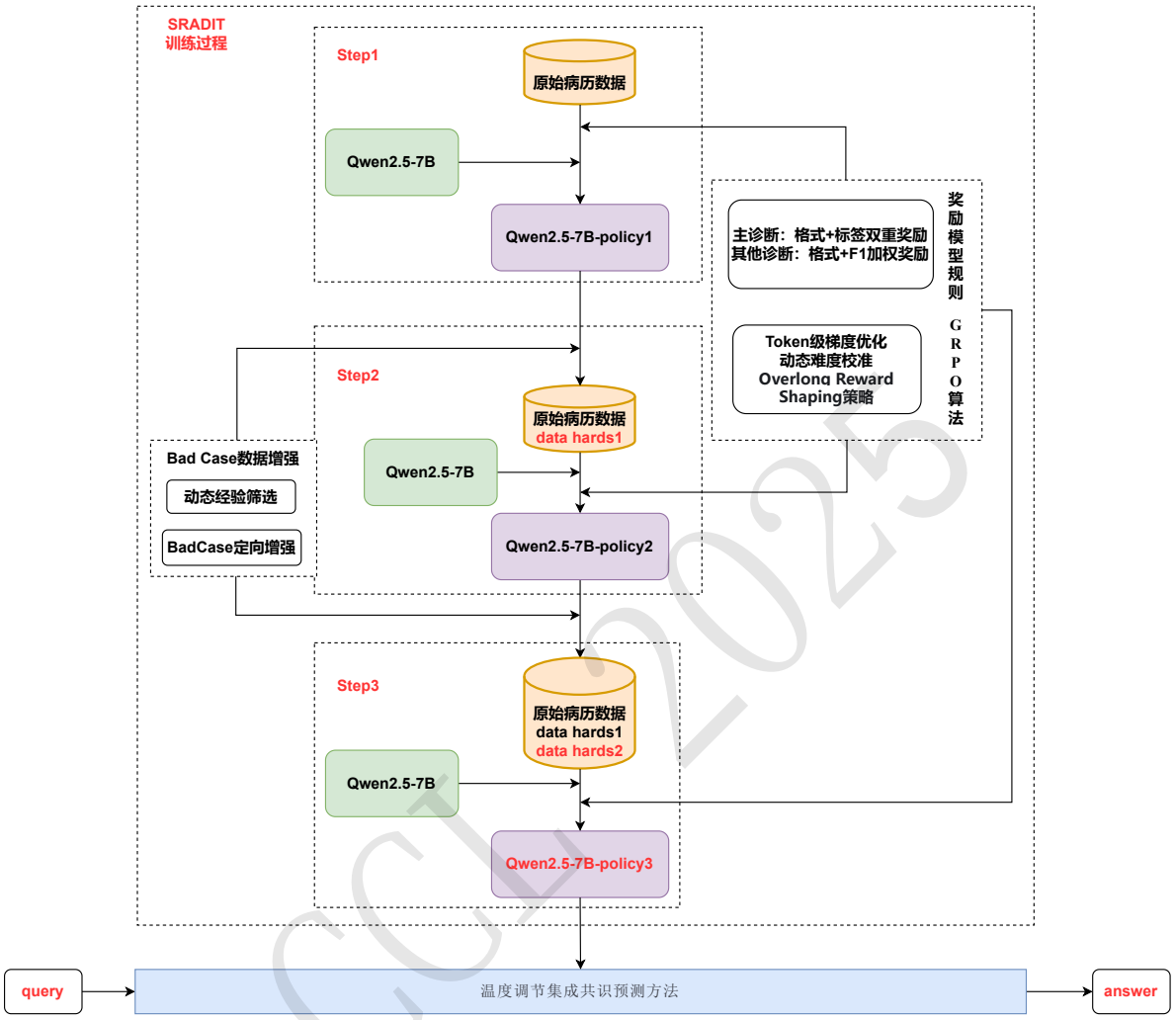

图 3: 基于规则奖励与自主思考强化学习的流程图

针对训练后期的策略优化，我们采用两阶段数据增强机制：

阶段一：动态经验筛选 当策略进入稳定区间（连续5个epoch的F1-Score波动<2%时），启动经验数据筛选机制；

阶段二：定向BDA 首先，基于稳定策略对训练集进行全量推理，筛选预测错误的样本构成Badcase集合$\mathcal{B}$。接着，采用分层改写策略：

1. 使用LLM对每个病历$d_i \in \mathcal{B}$进行语义保持改写，生成30个变体$d_i^{(j)} j = 1^{30}$，约束条件如公式（5）所示；

$$P(y|d_i^{(j)}) = P(y|d_i),\ \forall j \in [1, 30] \tag{5}$$

2. 通过SFT模型初筛和策略模型验证的双重过滤机制，分别剔除标签不一致的候选样本，以及剔除经过推理之后标签不一致的候选样本；

3. 人工专家从剩余候选集中选择语义最贴近原意的样本d_i^*加入增强集$\mathcal{B}^+$。

最后，将增强集$\mathcal{B}^+$合并至训练数据集，重新初始化策略进行训练。该机制通过“识别-生成-验证-增强”的闭环流程，进行三次迭代训练，显著提升模型在困难样本上的泛化能力。我们将这种方法称为策略轮动数据增强迭代训练（Strategy Rotation Augmented Data Iteration Train，SRADIT）方法，如图3所示。

2.5 温度调节集成共识预测方法

对于主诊断子任务，我们采用基于温度参数调节的生成式投票共识机制。具体步骤如下：

确定性生成： 设置温度参数$T = 0$，通过微调后的LLM生成首个类别预测结果$L1$，该设置确保输出具有最高确定性；

随机性生成： 调整温度参数$T = 0.7$，生成第二个类别预测结果$L2$，此设置允许模型引入适度随机性以探索不同输出空间；

共识决策： 若$L1 = L2$，直接将该类别作为最终预测；若$L1 != L2$，再次以$T = 0.7$生成第三个预测$L3$，通过多数投票机制确定结果。若存在出现次数大于等于2的类别，选择该类别；若$L1$，$L2$，$L3$互异，则采用$T = 0$时的确定性输出$L1$作为最终预测。

对于其他诊断子任务，我们采用基于温度调节的多轮生成共识筛选方法，具体流程如下：

多温度生成： 首先以$T = 0$生成初始标签集合$S1$，获取高确定性标签；随后两次以$T = 0.7$生成标签集合$S2$和$S3$，引入随机扰动以捕获潜在标签；

共识筛选： 对三次生成的标签集合进行频次统计，将在至少两次生成中出现的标签纳入最终预测集合S_{final}，构成最终结果。

该方法通过差异化温度设置平衡模型输出的确定性与多样性，利用集成共识机制提升多分类及多标签任务的预测可靠性，为大模型在医疗诊断等领域的精准推理提供有效解决方案。

3 实验

我们选择Qwen2.5-7B作为基座模型，使用第2节中我们设计的多个LLM微调策略，通过实验对比来验证方法在提升LLM中文电子病历ICD诊断自动编码任务的效果上的有效性。

3.1 数据集与评价指标

我们使用基于医院脱敏病历数据，其中训练集800条，A榜测试集200条，B榜测试集485条。训练集病历数据字段包括：病案标识、主述、现病史、既往史、个人史、婚姻史、家族史、入院情况、入院诊断、诊疗经过、出院情况、出院医嘱。数据标签分别为：主诊断编码、其他诊断编码。评价指标为:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\{0.5 \cdot I(y_{main} = \hat{y}_{main}) + 0.5 \cdot F1(y_{other}, \hat{y}_{other})\}_i \tag{6}$$

上述公式中，$I(\cdot)$为指示函数，满足条件返回1，否则返回0，$\hat{y}_{main}$和y_{main}分别表示主诊断编码的预测标签和真实标签；$\hat{y}_{other}$和y_{other}分别表示其他诊断编码的预测标签集和真实标签集；N表示测试样本数量；F1表示F1-Score值。

3.2 微调实验

3.2.1 实验设置

name	value
batch_size	32
learning_rate	5e-6
epoch	6
weight_decay	0.01
max_seq_len	8400
neftune_noise_alpha	5
flash_attn	fa2
warmup_ratio	0.1

表 1: SFT微调实验参数

本次实验我们使用的编程语言为Python，深度学习框架为PyTorch，训练框架为LLaMA-Factory，并使用DeepSpeed ZeRO-3分布式训练优化技术和全量微调技术。SFT微调实验的重要参数设置如表1所示。

对于强化微调，我们使用verl (Sheng et al. ，2024)训练框架，通过完全分片数据并行（Fully Sharded Data Parallel，FSDP） (Zhao et al. ，2023)训练加速，以及vLLM Rollout加速，其重要参数设置如表2所示。

name	value
train_batch_size	64
actor_lr	5e-7
ppo_epochs	1
ppo_mini_batch_size	16
ppo_micro_batch_size_per_gpu	2
group_num	6
adv_estimator	grpo
max_prompt_length	6350
max_response_length	1500

表 2: 强化微调实验参数

3.2.2 实验分析

基于第2节中提出的生成式微调方法、判别式微调方法，以及强化微调方法，微调后的Qwen2.5-7B模型在A榜和B榜测试集上的实验结果如表3所示。

微调策略方法	A榜	B榜
多任务MTL	78.5	\
子任务SFT	79.6	81.39
子任务CoTD-SFT	72.8	\
子任务K-Fold SFT	80.35	81.78
判别式微调	78.9	81.4
RBR偏好强化微调	72	\
RBR-R1强化微调-GRPO	79.8	\
RBR-R1强化微调-优化GRPO	80.78	82.18
RBR-R1强化微调-优化GRPO+SRADIT	**80.98**	**82.33**

表 3: 不同微调方法实验结果

由实验可以看出，生成式微调方法的效果要优于判别式微调方法。从实验结果来看，不带CoT的SFT效果反而优于带CoT的SFT，可能原因是带CoT的微调难度较大，小参数的LLM难以学会。在生成式微调方法中，K-Fold的SFT策略可以稳定取得很好的效果。

值得注意的是，强化微调在该任务上具有巨大的潜力，RBR-R1式带Think的强化微调方法在A榜和B榜上效果均优于其他微调方式，并且经过SRADIT的训练，效果得到了进一步提升。因此如果能够让模型探索学习到一个优质的策略，效果可以超过非强化微调类方法，充分利用困难样本进行增强训练，能达到更高的性能上限。

4 总结

本技术报告系统探索了LLM在中文电子病历ICD诊断编码自动生成任务上的能力，提出了生成式微调、判别式微调与强化学习微调的三重技术范式，并基于Qwen2.5-7B进行了详细的多策略微调实验对比。通过强化微调方法，我们显著提升了主诊断单标签预测的准确性与其他诊断多标签预测的完整性。实验表明，基于RBR-R1强化微调方法有效引导模型实现临床推理与

编码规范的对齐，在A/B榜评测中均突破传统监督微调的性能上限，验证了强化学习在医疗长文本复杂决策任务中的独特优势。

本技术报告仍存在一定局限性：其一，本次任务训练数据规模受限可能影响模型对长尾编码的泛化能力；其二，强化学习的训练效率与稳定性仍需进一步优化。后续工作将聚焦于多模态病历信息的联合编码建模、基于检索增强的少样本学习机制设计，以及轻量化部署方案探索，以推动医疗诊断系统在临床场景中的实际落地。

综上所述，本研究为LLM在医疗文本结构化任务中的应用提供了新的技术思路与实践参考，助力医疗信息化向智能化方向演进。

参考文献

Biyang Guo, He Wang, Wenyilin Xiao, Hong Chen, Zhuxin Lee, Songqiao Han, and Hailiang Huang. 2024. *Sample Design Engineering: An Empirical Study of What Makes Good Downstream Fine-Tuning Samples for LLMs.* arXiv preprint arXiv:2404.13033.

Chaojian Yu and Bo Han and Mingming Gong and Li Shen and Shiming Ge and Bo Du and Tongliang Liu. 2022. *Robust Weight Perturbation for Adversarial Training.* https://arxiv.org/abs/2205.14826.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, Chuan Wu. 2024. *HybridFlow: A Flexible and Efficient RLHF Framework.* Proceedings of the Twentieth European Conference on Computer Systems.

Hu, Jian and Liu, Jason Klein and Shen, Wei. 2025. *REINFORCE++: An Efficient RLHF Algorithm with Robustness to Both Prompt and Reward Models.* 2025REINFORCE.

Pavel Izmailov and Dmitrii Podoprikhin and Timur Garipov and Dmitry Vetrov and Andrew Gordon Wilson. 2025. *Averaging Weights Leads to Wider Optima and Better Generalization.* https://arxiv.org/abs/1803.05407.

Su, Jianlin and Zhu, Mingren and Murtadha, Ahmed and Pan, Shengfeng and Wen, Bo and Liu, Yunfeng. 2022. *ZLPR: A Novel Loss for Multi-label Classification.* arXiv preprint arXiv:2208.02955.

Shao, Zhihong and Wang, Peiyi and Zhu, Qihao and Xu, Runxin and Song, Junxiao and Bi, Xiao and Zhang, Haowei and Zhang, Mingchuan and Li, Y. K. and Wu, Y. 2024. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models.* https://arxiv.org/abs/2402.03300.

Teng, Fei and Liu, Yiming and Li, Tianrui and Zhang, Yi and Li, Shuangqing and Zhao, Yue. 2023. *A Review on Deep Neural Networks for ICD Coding.* IEEE Transactions on Knowledge and Data Engineering.

Wei, Jason and Wang, Xuezhi and Schuurmans, Dale and Bosma, Maarten and Ichter, Brian and Xia, Fei and Chi, Ed H. and Le, Quoc V. and Zhou, Denny. 2022. *Chain-of-thought prompting elicits reasoning in large language models.* Proceedings of the 36th International Conference on Neural Information Processing Systems.

Yanli Zhao and Andrew Gu and Rohan Varma and Liang Luo and Chien-Chin Huang and Min Xu and Less Wright and Hamid Shojanazeri and Myle Ott and Sam Shleifer and Alban Desmaison and Can Balioglu and Pritam Damania and Bernard Nguyen and Geeta Chauhan and Yuchen Hao and Ajit Mathews and Shen Li. 2023. *PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel.* IEEE Transactions on Computer Science.