

基于大语言模型的实验室报告解读系统构建研究 *

陆小琴^{1,2,3} 伍 柯² 雷玉倩² 唐 胡² 武宇翔⁴ 武永康^{1,2} 王 莉^{1,5}

(¹ 四川大学华西医院实验医学科 成都 610041 ² 金堂县第一人民医院信息与互联网医院管理部 成都 610400

³ 雅安职业技术学院药学与检验学院 雅安 625000 ⁴ 海南医科大学西英格兰学院 海口 571199

⁵ 四川大学华西天府医院妇科 成都 610213)

〔摘要〕 目的/意义 构建智能实验室报告解读系统, 帮助患者了解自身健康状况。方法/过程 收集四川省金堂县第一人民医院实验室检验报告并标注, 以 ERNIE-4.0-Turbo-8K 为基础模型, 用标注数据监督微调, 引入检索增强生成机制优化模型性能。结果/结论 监督微调后, 模型在实验室报告解读任务中多项评估指标优于微调前; 以项目编码为索引的检索增强生成机制, 能精准检索匹配知识切片, 提升解读准确性与可解释性。基于大语言模型优化的智能系统, 在提升医疗服务质量和效率方面潜力巨大, 但准确性、稳定性尚待深入评估。

〔关键词〕 大语言模型; 实验室报告解读; 监督微调; 检索增强生成

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.03.013

Study on the Construction of a Laboratory Report Interpretation System Based on Large Language Model

LU Xiaolin^{1,2,3}, WU Ke², LEI Yuqian², TANG Hu², WU Yuxiang⁴, WU Yongkang^{1,2}, WANG Li^{1,5}

¹ Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu 610041, China; ² Department of Information and Internet Hospital Management, Jintang First People's Hospital, Chengdu 610400, China; ³ School of Pharmacy and Laboratory Medicine, Ya'an Polytechnic College, Ya'an 625000, China; ⁴ UWE College of Hainan Medical University, Haikou 571199, China; ⁵ Department of Gynecology, West China Tianfu Hospital, Sichuan University, Chengdu 610213, China

〔Abstract〕 **Purpose/Significance** To build an intelligent laboratory report interpretation system to help patients understand their own health conditions. **Method/Process** The study collects and annotates the laboratory test reports from Jintang First People's Hospital in Sichuan province. Using ERNIE-4.0-Turbo-8K as the base model, it conducts supervised fine-tuning with the annotated data, and introduces the retrieval-augmented generation mechanism to optimize the model performance. **Result/Conclusion** After supervised fine-tuning, the model outperforms the pre-fine-tuning in multiple evaluation indicators for the laboratory report interpretation task. The retrieval-augmented generation mechanism indexed by item codes can accurately retrieve and match knowledge slices, improving the accuracy and interpretability of the interpretation. The optimized intelligent system based on large language models has great potential in improving the quality and efficiency of medical services. However, its accuracy and stability still need in-depth evaluation.

〔Keywords〕 large language models; interpretation of laboratory report; supervised fine-tuning; retrieval-augmented generation

〔修回日期〕 2025-01-20

〔作者简介〕 陆小琴, 副教授, 发表论文 11 篇; 通信作者: 王莉, 副主任医师。

〔基金项目〕 四川省科技计划项目 (项目编号: 2024YFFK0131)。

1 引言

随着现代医疗技术进步和健康意识提高，健康体检成为预防疾病和维护健康的重要手段^[1]。但体检报告中的专业术语和数据常使患者难以理解，尤其是实验室检验报告，往往以实验数据的形式呈现，需要医生解读。报告的专业解读是医院质量管理的重要内容^[2-3]。目前报告解读系统多基于规则方法，针对标准化指标效果较好，但缺乏灵活性，难以应对非典型数据和新问题，且为单一指标分析，易忽视指标关联性，无法提供针对个体的健康建议^[4-5]。因此，构建综合与个性化的智能实验室报告解读系统，既可帮助患者了解健康状况，又能减轻医生工作压力。

近年来，大语言模型（large language models, LLM）如 GPT、文心一言等，为实验室报告解读提供新方案^[6]。LLM 基于深度学习技术，通过海量文本数据无监督学习，生成具有强大自然语言处理能力的模型^[7]，能将复杂医学术语和数据转换为通俗易懂的解释。重要的是，这些模型可根据专业领域需求微调，以适应特定领域知识和语境，在特定任务上表现更出色^[8]。在生物医学数据上微调的通用领域 LLM 已取得令人印象深刻的结果。例如，基于 LLaMA^[9] 模型微调并使用 700 多种疾病医患对话数据训练的 ChatDoctor^[10]，药物推荐准确率达 91.25%；同样，从 LLaMA^[9] 模型微调而来的“华佗”^[11]，能提供中医药安全性和可用性建议。本研究探讨利用百度千帆大模型平台，基于 LLM 的监督微调（supervised fine-tuning, SFT）与检索增强生成（retrieval-augmented generation, RAG）技术构建智能实验室报告解读系统。

2 资料与方法

2.1 模型微调数据

2.1.1 数据来源 从四川省金堂县第一人民医院收集体检与门诊患者的实验室报告，每条数据代表一例患者的实验室报告。为确保模型训练效果，数据经过筛选、去重，保留 1 000 条用于模型微调。

患者年龄从 1 个月到 92 岁，其中男性 494 例，女性 506 例，数据包含 254 种异常项目类型，解读涉及生理因素和不同疾病导致的异常。所有数据均经脱敏处理，以保护患者隐私，防止信息泄漏。

2.1.2 数据标注 微调数据按输入（prompt）与输出（response）形式标注。经多次调试，设定 prompt 内容为：模型角色、解读要求、解读示例、患者情况、需解读的异常项目，示例如下。

#你是一个医学专家，需要你对患者的实验室检查结果进行解读。
#要求：

1. 首先对异常检验结果进行解读；
2. 综合所有检查项目进行解读（特别注意轻度异常的解读，要考虑生理因素或实验误差等因素，不要给患者造成心理压力）；
3. 给出进一步就医建议；
4. 给出健康建议；
5. 最后说明：以上内容为人工智能生成，若有任何不适或疑虑，建议及时就医咨询专业医生。

#解读示例：

<输入>：
男，43 岁，体检。乙型肝炎病毒表面抗原（HBsAg）45.269 IU/mL，参考范围 <0.05 IU/mL。

<输出>：
{
 "interpretations": [
 {
 "itemName": "乙型肝炎病毒表面抗原（HBsAg）",
 "checkResult": "45.269 IU/mL（升高），参考范围 <0.05 IU/mL",
 "itemDescription": ".....",
 "interpretation": "....."
 }
],
 "conclusion": ".....",
 "medicalAdvice": ".....",
 "healthAdvice": [
 ".....",
 "....."
],
 "notice": "以上内容为人工智能生成，仅供参考。若有任何不适或疑虑，建议及时就医咨询专业医生。"
}

#现在请你解读以下实验室检验结果：男，48 岁，体检。甘油三酯（TG）4.6 mmol/L，参考范围 0.4~1.7 mmol/L。

设定 response 内容为：异常项目解读、综合解读、就医建议和健康建议，输出文本采用 json 格式，便于提取关键词及构建展示格式，示例如下。

{
 "interpretations": [
 {
 "itemName": "甘油三酯（TG）",

```
"checkResult": "4.6 mmol/L (升高), 参考范围 0.4~1.7 mmol/L.",
"itemDescription": ".....",
"interpretation": "....."
},
],
"conclusion": ".....",
"medicalAdvice": ".....",
"healthAdvice": [
".....",
"....."
],
"notice": "以上内容为人工智能生成, 仅供参考。若有任何不适或疑虑, 建议及时就医咨询专业医生。"
```

2.1.3 数据划分 为科学评估模型的性能, 将标注数据集划分为 90% 训练集和 10% 测试集, 并确保随机分配以避免偏差。

2.2 模型构建关键技术

2.2.1 SFT SFT 是在预训练语言模型基础上, 使用标注数据进行微调的技术^[12]。通过在特定任务数据上训练模型参数, 优化其执行特定指令或任务的能力。SFT 适用于高度定制化场景, 能在保留预训练模型原有知识的同时, 以较少数据和计算资源实现任务优化。

2.2.2 RAG (图 1)

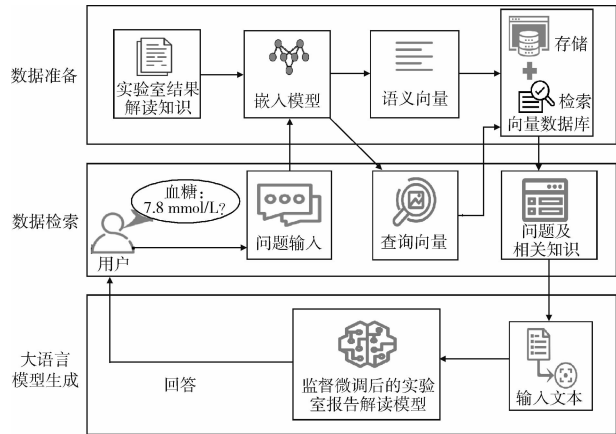


图 1 实验室报告解读模型 RAG 架构

RAG 结合检索与生成, 提高模型输出的准确性和可解释性。其从大型知识库检索与输入问题相关的信息, 作为上下文与输入一起处理, 使模型以较低成本获取专业领域知识, 生成可靠回答^[13-14]。RAG 结合检索和生成的优势, 使模型生成专业内

容, 并引用最新知识库内容, 提供权威且实时的信息。

2.3 实验室报告解读模型构建

2.3.1 模型构建平台 在百度智能云千帆大模型平台 (以下简称平台) 上对模型进行微调, 在其 Ap-pBuilder 平台中为微调后的模型引入 RAG 机制以优化模型性能。模型构建流程, 见图 2。

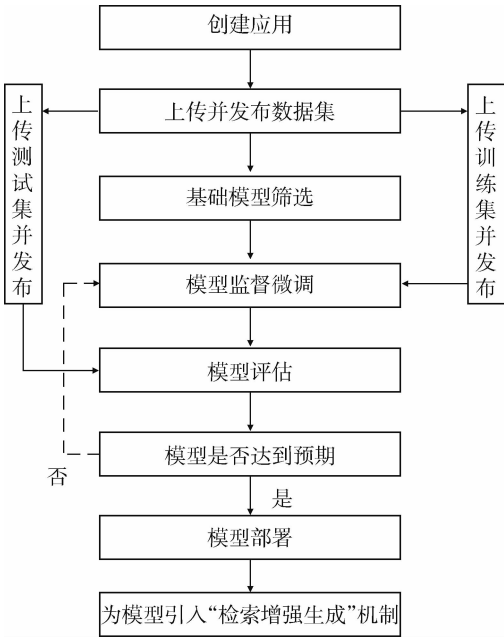


图 2 实验室报告解读模型构建流程

2.3.2 模型微调 在进行模型微调时, 首先要在平台控制台“应用接入”模块点击“创建应用”, 生成用于应用程序编程接口 (application program-ming interface, API) 调用的“API 密钥”(API Key) 和“秘密密钥”(Secret Key), 其中 API Key 标识应用身份, Secret Key 提供安全验证。接着在平台控制台选择“通用数据集”, 点击“创建数据集”上传训练与测试数据并发布, 为模型微调与评估做准备。然后在平台控制台“体验中心”试用各类基础模型, 依据试用效果及前期研究结果^[15], 选取 ERNIE-4.0-Turbo-8K 作为微调基础模型。随后在平台控制台选择“SFT”, 点击“创建训练作业”, 采用低秩自适应 (LoRA) 方法进行模型 SFT 调优, 核心训练参数为迭代 5 轮, 学习率 0.000 000 3, 序列长度 8 192, 全局批处理大小 18。完成调优后,

在平台控制台“模型评估”中，点击“创建评估任务”，选择测试数据集，以微调的实验室报告解读模型为基准，与 ERNIE-4.0-Turbo-8K 比对评估优劣。最后在平台控制台选择“我的模型”，点击“部署”，设置服务名称、模型 API 地址、付费方式等，部署在私有池，部署后外部系统便可通过 API 调用模型。

2.3.3 为模型引入 RAG 机制 在千帆 AppBuilder 平台“知识库”界面点击“创建知识库”，上传准备好的知识文本。知识文本采用 Excel 表格构建，表格设两列：第 1 列的列名为编码，其内容为检验项目的 ID 号与医保编号；第 2 列为检验项目临床意义等解读知识，将第 1 列编码设置为索引。设置文本索引配置后点击确定，平台对上传文本进行数据提取、文本分割、矢量化、数据入库处理，构建可检索的矢量数据库。上传入库后，生成检验项目知识切片列表，一行为一个切片，一个切片包含一个检验项目的编码与解读知识。在千帆 AppBuilder 平台的个人空间创建应用，将数据标注中的 prompt 设置为模型指令，部署的模型为问答模型，关联知识库，设置检索方法为关键字与语义混合检索模式。关联后，模型以编码为索引检索知识切片生成回答。

2.4 实验室报告解读系统构建

系统由模型、后端程序、前端界面构成。模型为关联知识库的实验室报告解读模型。后端程序连接前端与模型。前端界面搭建在金堂县第一人民医院互联网医院，患者点击报告解读时，后端程序提取异常检验结果及患者性别、年龄、诊断信息组成输入字段输入模型。输入包含子项 ID 及医保编码，触发 RAG 机制，模型在知识库中检索匹配切片，整合切片知识输出解读内容。内容传输到后端程序存储，医生审核后，后端程序美化处理 json 格式内容，再传输到前

端展示给患者。系统构建思路，见图 3。

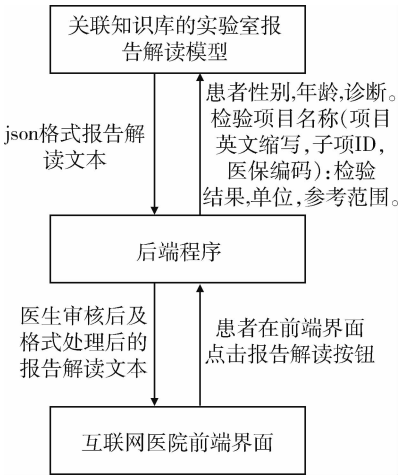


图 3 实验室报告解读系统构建思路

3 结果

3.1 SFT 微调结果

模型微调过程中的损失变化情况，见图 4。随着模型训练的逐步推进，其预测结果在训练集上的损失值持续降低，并逐渐趋近于 0，最终达到收敛状态，这表明模型在训练数据上的拟合效果不断优化，逐渐学习到了数据中的关键特征和模式。

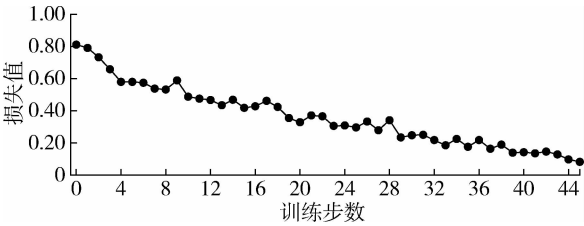


图 4 SFT 训练集损失曲线

3.2 模型评估结果

各评价指标结果，见表 1。

表 1 模型评估结果

模型	准确率	F1 (%)	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)	BLEU-4 (%)	MAUVE
ERNIE-4.0-Turbo-8K	0	64.29	64.34	42.13	57.24	33.58	0.93
实验室报告解读模型 (微调后的 ERNIE-4.0-Turbo-8K)	0	72.85	64.10	42.61	57.34	47.39	1.00

除准确率与 ROUGE -1 外, 其他评估指标微调后的模型均优于微调前的模型。

3.3 RAG 结果

微调后的模型关联知识库, 输入“患者性别, 年龄, 诊断. 检验项目名称 (项目英文缩写, 子项 ID, 项目医保编码): 检验结果, 单位, 参考范围”。知识库嵌入模型能根据输入编码精确检索完全匹配的知识切片, 并传输给解读模型, 检索匹配率 100%。解读模型根据专业切片知识进行解读, 有效提升解读的准确性和可解释性。

4 讨论

4.1 SFT 与 RAG 技术结合的优势

LLM 是基于深度学习的自然语言处理技术, 核心是 Transformer 架构。其通过大规模文本数据预训练学习语言结构和模式。文本生成过程中, 模型利用自注意力机制理解上下文, 逐步预测生成下一个词, 形成连贯高质量文本。LLM 技术在生物医学领域应用于生物医学信息检索、问答系统、文本摘要、信息提取和医学教育等^[16]。本研究利用预训练大语言模型, 结合监督微调和检索增强生成技术, 构建智能实验室报告解读模型及系统。模型评估结果显示, 微调后模型多项指标优于原模型。微调后 F1 分数从 64.29% 提升至 72.85%, 查准查全表现更佳。ROUGE 指标中, 原模型 ROUGE -1 略高于微调后, 但微调后 ROUGE -2 和 ROUGE -L 均高于原模型, 双字组重叠和最长相似片段匹配更优。BLEU -4 指标有所提升, 输出内容表述风格和准确性更佳。MAUVE 值从 0.93 提升到 1, 微调后模型语义理解和贴合度更佳, 能更准确解读实验室报告内容。综上, 微调后模型虽 ROUGE -1 稍弱, 但 F1、ROUGE -2、ROUGE -L、BLEU -4 和 MAUVE 等指标占优, 更适合实验室报告解读任务。模型评估微调前后准确率均为 0, 这是因为准确率评估要求模型输出与标注数据完全一致, 即便模型理解到位、解读方向正确, 文字表述不完全契合标注“标准答案”即判为错误。

实验室报告解读需最新准确知识, 但训练过程中模型学习到的知识可能存在偏差或不完整, 生成报告解读可能与事实不符, 即“幻觉”问题^[17]。研究^[18]表明 RAG 机制可减少知识缺乏导致的“幻觉”, 因此本研究为微调后模型引入 RAG 机制。实验室报告解读常涉及多个检验项目同时输入, 模型需准确检索匹配多个知识切片, 是研究难点。鉴于检验项目可从医院信息系统获取医保编码作为标识, 且检验项目的医保编码具有全国通用性, 将其加入 RAG 索引, 可增强知识库可推广性。但部分检验项目无独立医保编码, 因此将检验子项 ID 与医保编码组合成兼具唯一性与可推广性的项目标识, 将项目标识构建到知识库切片中并设为索引。同时设置后端程序, 从医院信息系统获取子项 ID 和医保编码并添加到输入字段。由此, 嵌入模型能根据输入项目编码, 100% 精确检索到完全匹配的多个切片, 有效解决难点, 避免检索不相关切片影响模型解读及占用输入词元 (tokens)。

综合利用 SFT 与 RAG 技术优化大语言模型, 实验室报告解读模型可提供更全面可信的解读服务, 有望提升用户体验和信任度, 但具体应用效果有待进一步评估。

4.2 系统实际应用面临的挑战

基于 LLM 优化的实验室报告解读系统能快速生成全面个性化解读报告, 但实际应用中仍面临挑战。首先, 需保障系统性能与稳定性, 尤其是在处理大量检验报告时, 系统应经全面测试以确保稳定运行。其次, 尽管引入 RAG 机制增强模型生成内容准确性和可解释性, 但模型生成基于概率和模式识别, 并非真正理解和推理^[16], 仍可能生成看似合理实则不正确的回答。因此系统后端设计审核机制中, 医生审核后解读报告才能展示给患者, 确保准确性。

4.3 本研究的局限性

本研究训练数据来源于单一医院, 数量有限, 可能限制模型泛化能力, 影响系统推广到其他医疗机构使用。此外, 实验室报告异常指标可能由多种

因素引起,模型需综合考虑患者其他检查结果才能提供更准确、全面的解读和建议,而目前训练数据仅包含实验室结果解读,未纳入影像学等检查结果。未来将进一步扩大和优化训练数据与知识库,提升模型泛化能力和实用性,为患者提供更优质的解读服务。

5 结语

基于 LLM 优化的智能实验室报告解读系统在提升医疗服务质量和效率方面具有潜力,但应用过程中需克服训练数据质量和多样性对模型性能及应用外推性的影响,系统输出准确性与稳定性需进一步评估与优化。未来将聚焦上述问题展开进一步研究,相信随着技术发展和研究深入,基于 LLM 技术的检验报告解读系统将在医疗健康领域发挥更加积极的作用。

作者贡献: 陆小琴负责数据标注、知识库构建、模型微调、论文撰写;伍柯负责原始数据下载与清洗;雷玉倩、唐胡、武宇翔负责数据标注;武永康负责研究设计、提供指导、知识库构建、论文审核;王莉负责标注数据审核。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 刘晓芸,王燕,邹媛. 基于个性化服务的健康管理模式在体检中心的应用及效果评估分析 [J]. 河北医药, 2024, 46 (13): 2024 - 2026.
- 2 范超,鲁卫平,黄庆,等. 以临床思维为导向的检验医师规范化培训模式探讨 [J]. 国际检验医学杂志, 2021, 42 (23): 2938 - 2939, 2944.
- 3 谭琪,范列英,虞珊珊,等. 以病例为基础的检验报告解读在检验医学科住院医师临床思维培养中的应用 [J]. 中国毕业后医学教育, 2024, 8 (10): 749 - 753.
- 4 胡占生,李小华,刘晓辉,等. 基于本体和 SWRL 规则分析健康体检结果的研究 [J]. 中国医学物理学杂志, 2014, 31 (3): 4901 - 4904, 4910.
- 5 柯娟玉,于帆,黄良娟,等. 微信端智能解读检验报告单系统的构建及应用 [J]. 海南医学, 2020, 31 (8): 1029 - 1032.
- 6 THIRUNAVUKARASU A J, TING D S J, ELANGO VAN K, et al. Large language models in medicine [J]. Nature medicine, 2023, 29 (8): 1930 - 1940.
- 7 柯沛,雷文强,黄民烈. 以 ChatGPT 为代表的大型语言模型研究进展 [J]. 中国科学基金, 2023, 37 (5): 714 - 723.
- 8 张钦彤,王昱超,王鹤羲,等. 大语言模型微调技术的研究综述 [J]. 计算机工程与应用, 2024, 60 (17): 17 - 33.
- 9 TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models [EB/OL]. [2024 - 02 - 27]. <https://arxiv.org/abs/2302.13971v1>.
- 10 LI Y, LI Z, ZHANG K, et al. ChatDoctor: a medical chat model fine - tuned on a large language model meta - AI (LLaMA) using medical domain knowledge [J]. Cureus, 2023, 15 (6): e40895.
- 11 WANG H C, LIU C, XI N, et al. HuaTuo: tuning LLaMA model with Chinese medical knowledge [EB/OL]. [2024 - 04 - 14]. <https://arxiv.org/abs/2304.06975v1>.
- 12 夏润泽,李丕绩. ChatGPT 大模型技术发展和应用 [J]. 数据采集与处理, 2023, 38 (5): 1017 - 1034.
- 13 LIU S, MCCOY A B, WRIGHT A. Improving large language model applications in biomedicine with retrieval - augmented generation: a systematic review, meta - analysis, and clinical development guidelines [J]. Journal of the American medical informatics association, 2025 (1): 1 - 11.
- 14 CALVIN W, JOSHUA O, CHARA W, et al. Potential for GPT technology to optimize future clinical decision - making using retrieval - augmented generation [J]. Annals of biomedical engineering, 2023, 52 (5): 1115 - 1118.
- 15 陆小琴,佳薇,武宇翔,等. 大语言模型在检验医学领域的应用潜力与挑战评估 [J]. 临床检验杂志, 2024, 42 (8): 619 - 623.
- 16 TIAN S, JIN Q, YEGANOVA L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health [J]. Briefings in bioinformatics, 2023, 25 (1): 1 - 13.
- 17 颜见智,何雨鑫,骆子烨,等. 生成式大语言模型在医疗领域的潜在典型应用与面临的挑战 [J]. 医学信息学杂志, 2023, 44 (9): 23 - 31.
- 18 GUO Y, QIU W, LEROY G, et al. Retrieval augmentation of large language models for lay language generation [J]. Journal of biomedical informatics, 2024, 149 (1): 104580.