# Project 2：医学语言生成实验报告

## 1. 实验任务与数据集

本实验选用 MedBench 框架下的 IMCS-V2-MRG（Medical Report Generation） 数据集。任务目标是：根据中文门诊多轮医患对话文本自动生成结构化医疗报告，包括以下六个标准字段：主诉、现病史、辅助检查、既往史、诊断、建议。  
  
报告生成结果需符合格式约束（严格六段式输出），并以字级 ROUGE-1 / ROUGE-2 / ROUGE-L 为主要评价指标。最终指标取三者算术平均。

## 2. 模型与环境配置

本实验在本地多 GPU 环境（NVIDIA RTX 3090×8）上部署两个不同结构的大语言模型（LLMs）：Qwen2.5-7B-Instruct 和 Llama-3.1-8B-Instruct。两者均使用 vLLM 框架通过 OpenAI API 接口调用，并统一参数：max\_tokens=512, temperature=0.2。  
  
依赖环境包括 Python 3.10、vllm、openai、rouge、jsonlines、argparse。辅助脚本 convert\_for\_eval.py 用于格式转换，eval.py 负责计算 ROUGE 得分。

## 3. 实验流程

（1）数据准备：加载 IMCS-V2-MRG 测试集并规范化。  
（2）推理生成：分别运行 Qwen 和 Llama 模型生成预测报告。  
（3）结果转换：调用 convert\_for\_eval.py 将 JSONL 文件转为评测格式。  
（4）性能评测：通过 eval.py 计算字级 ROUGE-1/2/L。

## 4. 实验结果

评测结果如下：  
  
模型 ROUGE-1 ROUGE-2 ROUGE-L 平均值  
Qwen2.5-7B-Instruct 0.4620 0.2803 0.3883 0.3769  
Llama-3.1-8B-Instruct 0.4382 0.2517 0.3752 0.3550  
  
Qwen 模型在所有指标上均优于 Llama，平均 ROUGE 提升约 6.2%。特别是在 ROUGE-2（短语级复现）指标上提升显著，显示其在中文生成上的优势。

## 5. 分析与讨论

1. 六段式模板有效约束输出结构，提升格式一致性；  
2. Qwen 针对中文优化，因此在语义连贯和术语复现上更优；  
3. ROUGE 只能衡量字面相似度，未反映医学逻辑一致性；  
4. 平均推理延迟约 2.3 秒/样本，吞吐率约 0.4 req/s。

## 6. 结论与展望

结论：Qwen2.5-7B-Instruct 在 IMCS-V2-MRG 任务上取得最佳综合表现（平均 ROUGE=0.3769），Llama 表现略逊但仍具结构化能力。  
  
展望：  
- 引入多任务指令微调以增强泛化；  
- 尝试 Chain-of-Thought 或 schema-guided prompt 改善诊断一致性；  
- 利用生成报告进一步构建医疗知识图谱前端模块。

## 附录：文件结构

run\_infer\_imcs\_mrg.py：执行推理  
convert\_for\_eval.py：格式转换  
eval.py：ROUGE 计算  
results.txt：评测结果  
test.json：测试样例  
README.md：实验说明文档