黄景行

jxhuang2003@stu.xjtu.edu.cn · (+86) 15972407932 · **求职方向：llm 应用/AIGC**

教育背景

西安交通大学 2020.09 – 2025.06

本科 物联网工程

西安交通大学 2025.09 – 2028.06

硕士（保送）计算机科学与技术

实习经历

上海得物信息集团信息有限公司 2025 年 3 月 – 2025 年 6 月

- 使用 **langgraph** 构建 **agent**，使用提示词工程实现意图识别功能以及实现多轮对话功能
- 实现 agent 的 **text2sql** 功能：
  - 构建训练知识数据集：包括数据库中表的 DDL，document 以及 question-sql 问答对，使用该数据集训练 rag 系统。
  - 构建 **rag** 系统，使用 ONNXMiniLM_L6_V2 嵌入模型和 ChromaDB 向量数据库，根据用户的问题召回与问题最相关的 SQL、DDL 和文档
  - 构建动态 **prompt**，将 rag 系统召回的内容，以及通过 COT 的方法 llm 生成的 text2sql 的推理流程都放入上下文窗口，使准确性大幅提高约 36%。

项目经历

AI Mathematical Olympiad kaggle 2024 年 4 月 – 2024 年 7 月

NLP、*LLM* 团队项目

比赛地址：https://www.kaggle.com/competitions/ai-mathematical-olympiad-prize

项目背景

- **目标：**创建能够解决 LaTeX 格式的复杂数学问题的算法，提升 AI 模型的数学推理能力。具体来说，解决 110 个类似中级水平高中数学挑战的问题，答案是一个非负整数。
- **评估指标：**预测标签与真实标签的准确率。

解决方案和关键技术

- **数据集：**使用 AMC 数据集选择最新的 50 道题作为本地验证集。
- **模型选择：**采用 deepseek-math-7b-rl 模型，使用温度（temperature）为 0.9，top_p 为 1.0，最大 token 为 2048。在两张 T4 GPU 上并行运行模型，最大化自一致性以提升预测准确率。
- **时间管理：**创建了 TimeManager 类，动态调整每道题的尝试次数，确保在规定时间内尽可能多地解决问题。通过减少困难题目的尝试次数，提高其他题目的尝试次数。
- **候选答案生成与比较：**使用两个 prompt 分别通过 cot 和 write code 的方式生成候选答案。限定每个 repetition 只有 3 次修正代码的机会，防止输入文本过长导致输出质量下降。使用加权计数排序候选答案，减少模型对 0-5 等小数的权重。修复非有效答案（如 error、小数、复数等）时的处理逻辑，跳过无效答案记录。
- **优化双 GPU 运行：**使用 ThreadPoolExecutor 并行回答问题，提高双 GPU 的利用率，减少互相等待时间，提升整体 repetitions 数量。
- **日志记录：**详细记录每次实验的数据，包括每个问题的所有候选答案、program code，code accuracy，text accuracy 等，帮助快速实验和结果观察。

成果与效果数据

- **性能提升：**双 GPU 部署将每题的总 repetitions 数量从 21 次提升至 30-40 次。
- **准确率：**本地实验得分约为 21，pass1@ 达到 32，pass2@ 达到 23。
- **Public Leaderboard：**不同的 seed 下得分差异为 3 分，体现了方案的稳定性与有效性。

基于 RAG 的汽车知识问答系统 2024 年 9 月 – 2024 年 12 月

RAG, LLM 个人项目

项目地址：https://github.com/winter-JX/rag_with_chat

项目背景

本项目属于大模型 RAG 任务，使用现有的车主手册构建知识库，然后选择知识库中的相关知识用于辅助大模型生成。整个方案的构建流程主要分为三大部分：构建知识库、知识检索、答案生成。该项目主要结合了 LLM、Langchain、提示工程、优化知识库结构和检索生成流程、vllm 推理优化框架等技术。

解决方案和关键技术

- **数据集：**训练数据集主要是一本汽车的用户手册（pdf 文件）。
- **模型选择：**采用 Qwen2.5-7B-Instruct 大语言模型，bge-reranker-large 重排序模型，bge-m3 文本嵌入模型，text2vec-base-chinese 相似度模型。
- **pdf 解析：**对于 pdf 文件中的文本内容，采用了三种解析方案的综合。
  - **pdf 分块解析：**尽量保证一个小标题 + 对应文档在一个文档块，其中文档块的长度分别是 512 和 1024。
  - **pdf 滑窗法解析：**把文档句号分割，然后构建滑动窗口，其中文档块的长度分别是 256 和 512。
  - **pdf 非滑窗法解析：**把文档句号分割，然后按照文档块预设尺寸均匀切分，其中文档块的长度分别是 256 和 512。

  按照这个三种解析方案对数据处理之后，然后对文档块做了一个去重，最后把这些文档块输入给召回模块。使用三种解析方法的综合，可以保证文本内容的完整性和跨页连续性。
- **召回：**召回主要使用 langchain 中的 retrievers 进行文本的召回，选用了两种召回方法：
  - **深度语义召回：**使用了 m3e 召回和 bge 召回两种方法。
  - **字面召回：**使用了 BM25 召回和 TF-IDF 召回两种方法。
- **重排序：**分别使用了 bge-reranker 和 bce-reranker-base_v1 模型对检索召回的文档进行重排。
- **vllm 推理优化：**LLM 采用，Qwen2.5-7B-Instruct 作为大模型基座，并且都使用了 vllm 框架来进行加速推理优化。。

成果与效果数据相比原生 LLM 外挂知识库提升 4.1%。

IT 技能

- 编程语言: Python == C++ > C > 其他
- 框架和平台: pytorch、Linux、Windows
- 熟练掌握机器学习, 大模型和计算机基础相关知识，例如**数据处理，模型评估，transformer 架构，llama-factory 框架，RAG，Agent** 等

获奖情况

美国大学生数学建模大赛, 特等奖提名 (2.5%) 2024 年 5 月

- 减少非法野生生物贸易解决方案。数据挖掘、英文文献阅读与写作

kaggle：*AI Mathematical Olympiad - Progress Prize 1*, 金牌 (4/1161) 2024 年 7 月

- 微调大模型构建解答以 latex 语言表达的数学问题的解决方案。数据集构建，python 编程

全国大学生机械创新设计大赛全国二等奖, 慧鱼组全国一等奖（3%） 2024 年 4 月

- 设计一款智能莲藕采收机器人以实现多种功能。无线控制、嵌入式系统

全国大学生物联网设计大赛, 全国一等奖（3%） 2024 年 8 月

- 使用物联网，大模型等技术构建智慧水务解决方案。物联网系统、计算机视觉、大模型智能体

全国大学生数学建模大赛, 陕西省一等奖 2023 年 10 月
中国机器人及人工智能大赛智能物流组, 全国二等奖 2022 年 7 月
APMCM 亚太地区大学生数学建模竞赛, 二等奖 两次
国家级大创，省级大创，校级大创结项 各一项

以上比赛均为第一负责人，以及校级奖项若干

其他

- 语言: 英语 - 熟练 (通过 CET6)
- 性格：乐观开朗，乐于团队合作，擅长沟通交流，对大模型相关技术有热情，自驱力强，愿意积极探索