鲁国伟

23岁 | 男 | 湖南岳阳 | 2年经验
17673472715 | 3485135655@qq.com

项目经验

基于Qwen的智能纠错辅助系统 - **技术负责人** 2025-02 ~ 2025-06

Situation：针对学生使用编程学习平台时遇到的 “做题卡壳” 和 “纠错困难” 问题，结合大模型（Qwen）当前答案准确率低、操作复杂的情况，可以设计一套 分层辅助系统，通过 “轻量级人工干预+AI优化” 提高效率。

技术栈：Qwen模型 | VLLM加速推理 | PEFT微调框架 | accelerate分布式训练

Action：

- 数据集清洗：对于数据库中的题目使用Qwen大模型提取关键性信息并存储成结构化数据
- 微调数据制作：人工标注+半自动化生成共9k条微调数据，语法错误3k+逻辑错误3k+不会开头3k
- 模型微调：微调Qwen2.5-7B意图识别模型（95%准确率），区别问题类型。结合**RAG技术**，获取数据库中的题目和答案，使用**PEFT-Lora框架微调**Qwen2.5-32B模型，并结合**accelerate框架实现多卡训练和混合精度训练（GPU显存减少70%）**。
- 评估指标：使用模型评估（BLEU-4、ROUGE-L）+人工评估+**分阶段评估（**先小规模AB测试，再全量推广）
- 模型量化：对Qwen2.5-32B使用 **GPTQ/AWQ** 量化（4-bit/8-bit），减少显存占用和计算量。对意图识别模型（Qwen2-7B）采用 **动态量化**（torch.quantization），提升CPU推理速度。
- 加速推理：使用**VLLM框架**，启用 **PagedAttention** 和 **Continuous Batching**，支持高并发。使用 **TensorRT-LLM** 部署量化后的模型。
- 部署：Docker封装全链路服务。

Result：

✓系统支持三种场景问题解答，5K测试集上答案生成准确率87%（LLMs:73%）

✓通过量化、缓存等技术，平均响应时间压降至2.7秒(首Token0.8s)，满足实时交互需求。

✓多模型协同部署方案节省2倍GPU资源，支持高并发场景下的稳定服务。

✓为深度学习课程作业提供高效辅助工具。

课业通鉴系统 - **技术负责人** 2024-07 ~ 2025-01

Situation：基于教师对不同的学生，需要根据课程进度和学生性格挑选不同的课程和比赛。针对这种情况需要构建一个高效、精准的自动化查询和问答系统，可以对每一个学生的情况进行个人的评估和课程的挑选，以及针对学生的进度和能力推荐比赛的等级和题目。这项任务对于教师过于繁琐，需要结合**大模型语言（LLM）与专业领域知识处理技术**，实现高准确率、低延迟的服务。

技术栈：PGSQL | 分块策略 | BM25 | 多路召回架构 | VLLM部署加速推理 |

Action：

- 数据解析与清洗：使用pdf工具（PyMuPDF）解析PDF文本，基于OCR工具（Tesseract OCR**）**解析图片型文档内容，提取关键元素；对于文本类文档， 通过大语言模型提取关键信息并存储为结构化数据。
- RAG知识库构建：对收集的数据进行设计**分块策略（递归分块/语义分割）**，并使用Qwen2.5-32B进行prompt提示提取文档元知识（课程名称/内容/知识点等），使用**嵌入模型**（**BGE-M3，相比OpenAI Embeddings平均F1增加3%**），构建关系数据库（pgsql）。
- RAG召回：微调Qwen2-7B模型（98%的落域准确率），区分任务类型和提取关键词。对数据库存储中的元知识向量进行余弦相似度计算，得到最优数据。
- RAG优化：使用混合召回框架，**BGE语义相似度+BM25关键字匹配**，通过**相似度Top20→关键字匹配Top1等**策略进行召回优化，知识命中率从85%→91%。
- 生成优化：实现**Query重写管道**（纠错/同意拓展/术语标准化），问答准确率+2%。
- 评估体系：采用**模型评估+人工评估**，自动化指标：BLEU-4、ROUGE-L、知识覆盖度。人工审核：对模型低置信度结果抽样复审，最终回答优秀率94%。
- 模型部署：基于**VLLM部署**Qwen2-7B。

Result：

✓ 测试集准确率94%。

✓ 支持日均3.2千次查询，API P99延迟<1.2s。

✓ 技术咨询重复问题减少67%，人工抽检满意度93%。

工作经验

广州小飞侠教育科技有限公司 - **大模型算法工程师** 2024-07 ~ 2025-06

1. 文档分块与向量化，设计分块策略，选择嵌入模型，构建向量数据库（pgsql）。
2. 检索算法优化，实现混合检索（关键词BM25+向量检索），调整相似度阈值，加入元数据过滤（如文档来源、时效性）。
3. 性能调优，量化嵌入模型（FP16→INT8），测试HNSW/PQ等索引算法。
4. 实现监督微调（SFT）、RLHF（PPO/DPO）或对比学习（CPT）
5. 调试LoRA/QLoRA/P-Tuning等参数高效微调方法
6. 实现量化（GPTQ/AWQ）、模型剪枝、知识蒸馏（如DistilBERT→TinyLLM）
7. 部署vLLM/TensorRT-LLM等推理框架，优化KV Cache

腾讯微保 - **大模型算法实习生** 2023-10 ~ 2024-01

1. 参与大规模语料清洗、数据预处理（去噪、去重、标准化）
2. 构建高质量的指令数据集（如Self-Instruct数据生成）
3. 调试LoRA/QLoRA/P-Tuning等参数高效微调方法

上海梦孚教育科技有限公司 - **NLP算法实习生** 2023-06 ~ 2023-08

1. 参与大规模语料清洗、数据预处理（去噪、去重、标准化）
2. 对Bret、、CILP等模型进行微调
3. 设计课程学习（Curriculum Learning）策略或数据采样策略
4. 实验跟踪：Loss曲线分析、训练效率优化（如梯度裁剪、学习率调度）

技能特长

- 熟练使用Python编程，熟悉Linux开发环境及Shell脚本，了解C/C++基础开发。
- 熟悉PyTorch框架，熟悉Hugging Face生态，掌握LangChain等大模型应用开发工具链。
- 熟悉大语言模型（LLM）核心技术，包括检索增强生成（RAG）、提示词工程（Prompt Engineering）、模型微调与对齐方法、Agent。
- 熟悉DeepSpeed分布式训练框架，具备多机多卡GPU集群训练经验。
- 掌握高效微调技术（LoRA、QLoRA、P-Tuning），熟悉LLaMA-Factory等轻量化训练框架。
- 熟悉大模型高性能推理框架（vLLM），具备模型服务API开发经验。
- 掌握Docker容器化部署，能独立完成模型封装、服务部署及运维。

教育背景

湖南农业大学 - **智能科学与技术** - **本科** 2020-09 ~ 2024-06

- 大学生创新创业项目负责人
- 人工智能实验室负责人