齐雨凡

应聘岗位：大模型算法工程师

杭州市钱塘区

15007210384

15007210384@163.com

教育背景

2017.9~2021.7	**武昌理工学院**	**计算机科学与技术**	**本科**

职业专长

◆数据处理技术：

- 擅长数据合成、数据增强与清洗，熟练运用 data-juice、self-instruct 等工具和技术思路，对数据进行预处理，并清理脏数据和对数据做优化，为模型训练提供高质量数据。通过数据合成和增强技术，扩充数据规模、增加数据多样性，有效提升模型的泛化能力和训练效果。

◆数据集制作：

- 精通 JSONL 格式的 sft 数据制作，涵盖指令微调数据和多轮对话数据。熟悉 Alpaca 格式的指令监督微调数据集和 sharegpt 格式的多轮对话 message 数据制作流程，可制作高质量数据集用于模型训练。

◆LLM 性能测试与调用:

- 具备丰富的 LLM 性能测试经验，能够运用科学的测试方法评估模型性能。通过调用 API，对 gemini-2.0-flash-exp、claude-3.5-sonnet-20241022、gpt-4o-2024-11-20、qwen-max-1201、qwen2-72b-instruct、deepseek-chat、deepseek-r1 等多种 LLM 进行性能测试，根据测试结果精准选择合适的基座进行训练，深入了解不同模型的优势与不足。

◆大语言模型与视觉语言模型应用及微调：

- 熟练使用 qwen2.5_14b-instruct、llama3.1-chinese、internlm2_5-7b-chat、deepseek-vl-7b-chat、qwen2.5_7b-instruct 等 LLM 和 VLM，并掌握其 fine tunes 技术。能够根据不同任务需求，对模型进行针对性微调，提高模型在特定领域的性能表现。

◆大模型精准调优效能评测框架：

- 熟悉 ms-swift、llama-factory 的 fine tunes 评估框架，能够运用其对模型微调效果进行有效评估，为模型优化提供数据支持和方向指引。

◆Fine tunes 技术:

- 掌握 sft、全量 fine tunes、lora 等 fine tunes 技术，熟悉相关参数设置。熟练运用分布式 train 技术，如 deepspeed zero2（可提高 train 速度）和 zero3（可减少显存但速度会变慢），提升模型训练效率和资源利用率。

◆高性能推理框架:

- 精通 xinference、vllm、ollama 的 inference 框架，熟练掌握其在模型推理阶段的部署与应用，确保模型高效、稳定地输出推理结果。

◆AI 应用开发框架:

- 熟练使用 langchain、langgraph、llamaindex、dify、fastgpt 框架，能够基于这些框架进行 RAG，Agent 项目的开发，实现如智能问答、文本生成等功能，提升项目开发效率和质量。

◆测试页面搭建:

- 熟练使用 gradio、streamlit 搭建测试页面，为测试人员提供便捷的测试环境，便于对模型和项目进行功能测试和效果评估。

◆**Prompt 工程:**

■ 擅长 prompt 的精心设计，包括人设设置、数据处理、数据生成等任务。通过优化 prompt，有效引导模型生成高质量文本，提升模型在各类场景下的应用效果。

◆**前沿技术探索与研习:**

■ 不断拓宽技术视野。通过学习最新资讯，如 deepseek r1 训练过程、ktransfermer、unsloth 微调框架、嵌入式 AI 小智的制作、grok3 等，保持对行业前沿技术的敏锐度。

◆**技术知识沉淀与传播:**

■ 在个人技术 blog 发布三篇技术文章，尽管内容偏基础，但通过写作与分享，不仅巩固了自身所学知识，也展现了对技术的总结和输出能力

■ 个人博客链接:个人博客

◆**AI 驱动的集成开发环境（IDE）高级插件体系运用:**

■ 熟练使用 vscode cursor，掌握其中 Copilot AI 插件的使用技巧，同时熟悉 trae cursor、windsurf 相关操作，能借助这些工具提升 AI 编程效率，实现代码快速生成、智能补全等功能。

职业经历：

2024.9-2025.1　　杭州焱黄智能科技有限公司　　NLP 算法工程师

- RAG 项目推进：独立开发 RAG 系统，达成多轮对话智能问答，依托知识库减少模型幻觉，提升垂直领域问答效果；负责 RAG 系统文本数据的分析、清洗与切分。
- 数据处理优化：独立完成 Fine tunes 数据合成、增强与清洗，提升数据多样性与整体质量；规划并处理情感对话数据，进行标注整理。
- LLM 全流程把控：独立负责 LLM 选型、训练、微调、性能优化、评估及高性能部署。多次微调模型并调整参数，用指标评估模型表现。
- 技术研究创新：研究 self-instruct 等数据合成思路、data-juice 算子并清理数据，学习 COSTAR 框架制作 Prompt，优化训练参数并解决技术问题。

2021.8-2024.8　　深圳市好邻养老科技有限公司　　NLP 算法工程师

- VLM 模型微调（2024.6-2024.8）：开展 VLM 模型微调工作，结合业务需求与数据特征，对模型进行针对性优化。提升了模型在特定视觉-语言任务上的性能表现，为业务应用提供支持。
- Agent 工具开发（2024.1-2024.6）：参与基于 Tavily 搜索 API 的 agent 实现工具开发调试，借助 langgraph、langchain 等库达成工具灵活调用与实时信息查询，从获取 API 密钥、初始化组件到设计工具调用逻辑、构建状态机，再到集成流式处理与 gradio 界面，经算法和架构优化，提升了 agent 在复杂环境下的响应及查询能力，增强了系统实用性与交互性。
- RAG 本地化开发部署（2023.7-2023.12）：运用 llamaindex 框架完成 rag 的本地化开发及部署。参与系统架构设计、数据处理、模型集成和优化等项目实施全流程，保障 rag 系统在本地环境稳定运行，满足业务知识检索和问答需求。
- 养老产品海报设计（2022.11-2023.6）：基于 SD 模型开展养老产品海报设计与生成工作。深入了解养老产品特点和目标受众需求，通过调整模型参数和输入条件，输出符合品牌形象和市场定位的高质量海报，提升产品宣传效果。
- 养老命名实体识别（2022.4-2022.10）：负责基于 Bert＋LSTM＋CRF 的养老命名实体识别项目。把控从数据收集、预处理到模型训练和评估的项目全流程。优化模型结构和参数，提高实体识别的准确率和召回率，为养老领域信息提取和分析提供工具。
- 养老数据文本纠错（2021.8-2022.3）：开展基于 T5 的养老数据文本纠错项目。利用 T5 模型文本生成能力，结合养老数据特点开发文本纠错系统。经大量实验和优化，降低养老数据文本错误率，提高数据质量和可用性。

项目经历：

2024.9-2024.10　　杭州焱黄智能科技有限公司　　AI 恋爱军师

- **项目背景：**

  在现代社会，人们对情感咨询的需求日益增长，但传统的情感导师服务面临着诸多难题。一方面，高昂的咨询费用让许多人望而却步；另一方面，由于情感导师数量有限，响应速度较慢，无法及时满足用户的需求。为了打破这些困境，我们团队全力打造了 “AI 恋爱军师” 项目。该项目依托微信公众号平台，借助大模型技术，为用户提供高效、个性化且随时随地可得的情感咨询服务，帮助用户快速解决情感困扰。

- **核心技术** ： Fastgpt,Xinference,OneAPI,Chatgpt-on-wechat

- **项目实施:**

  - **技术选型：**
    - 采用 Xinference 本地化部署 Qwen2.5 模型，通过 int4 量化降低显存消耗，支持高并发情感咨询场景；
    - 基于 FastGPT 搭建 RAG 系统，整合情感导师课程、书籍等结构化知识库，实现精准问答；
    - 通过 ChatGPT-on-Wechat 接入微信公众号，完成用户对话交互闭环。
  - **实现流程：**
    - 用户输入问题后，系统优先从预置话术库匹配高频问题（如“如何挽回感情”）；
    - 复杂问题触发 RAG 检索，从 PDF/Word/视频解析的知识库中提取关键建议；
    - 结合军师人设模板（如“朋友啊，这事儿得这么看…”）生成最终回复。

- **项目成果：**

  高效的情感咨询服务：成功实现 7x24 小时不间断实时情感咨询服务，用户借助微信公众号，随时都能获取情感建议。优化推理速度后，平均响应时间低于 2 秒，大大提升了服务效率，用户满意度达到 90% 以上，有效解决了传统服务响应慢的问题。

  高质量的内容生成: 借助优化后的 RAG 系统和混合检索技术，AI 恋爱军师能够为用户提供精准、实用的情感建议。在情感咨询场景下，对话准确率从 70% 左右提升至 85% 以上，切实帮助用户解决情感难题，提供可靠的情感解决方案。

  可扩展的技术架构：项目采用模块化设计理念，具备良好的扩展性和迭代能力。比如在模型服务层，可便捷地替换或升级推理引擎、Embedding 模型等组件。通过 OneAPI 统一管理接口，大幅降低了后续维护和开发的复杂度，为项目的持续发展和功能拓展提供有力保障。

2024.10-2025.1　　杭州焱黄智能科技有限公司　　海王角色扮演模型

- **项目背景：**

  在数字化社交盛行的时代，线上交流成为拓展社交圈和建立情感联系的重要方式，异性线上互动愈发频繁。有权威调研显示，18 - 35 岁年轻群体中，75% 的人在与异性线上聊天时会遇到回复难题，影响交流。
  男性在与女生线上交流时挑战重重，例如面对女生金钱相关的 “废物测试”，超 60% 的男性难以巧妙回应；面对女生间接性高冷，约 70% 的男性不知如何活跃气氛、延续话题，这导致他们自信心受挫，还可能错失情感发展机会。
  目前，市面上的交流辅助工具多为通用建议，无法结合具体情境精准回复。现有语言模型处理情感交流场景时，难以捕捉情感信号和潜在含义，回复生硬、缺乏共鸣，无法满足真实社交需求。
  为此，“海王角色模型” 项目启动。项目团队利用先进的自然语言处理技术，收集大量异性情感交流数据并通过合成数据的手段增加数据的多样性，深度微调基础模型。目标是打造高情商、风趣幽默、能捕捉情感信号的海王风格模型，帮助男性在与女生线上交流时准确理解对方意图，给出有吸引力的回复，增强自信，提升情感交流质量与成功率，填补市场空白。

- **项目技术：**数据合成，Data-juicer,Ms-swift,Vllm,Dify,qwen2.5-14b-instruct
- **项目实施：**
  - **模型选型与数据准备：**在深入调研基座模型后，选定 qwen2.5 - 14b - instruct 模型作为微调基础。对人工收集的异性情感交流对话数据进行全面分析，运用 Data - juicer 框架执行数据清洗任务，去除重复、低质量数据，提升数据整体质量。同时，借助 AI 技术实现自动化打标签，将原始数据转化为适用于 sft 微调的格式。
  - **模型微调：**采用 Ms - swift 微调框架，运用 sft、lora 方法对选定模型进行微调。在多次微调过程中，密切关注模型表现，发现效果未达预期。经过深入分析，决定调用 Gemini2.0 - flash - exp 模型的 api，对数据进行垂直领域幽默海王角色风格和女多轮对话合成。合成后的数据通过数据扩充手段增加数据量，利用数据增强技术提升数据多样性。之后再次使用 Data - juicer 进行数据清洗，获得高质量数据，为后续微调提供有力支持。
  - **模型部署与应用开发：**利用 Vllm 进行模型部署，搭建起稳定的服务环境，确保模型能够高效运行并响应请求。运用 Dify 进行 AI 应用开发，将微调后的海王角色模型融入其中进行内部测试。
- **项目成果：**
  - **模型性能提升：**成功打造出高情商、风趣幽默且善于捕捉情感信号的海王风格模型。该模型在面对女生金钱相关的 “废物测试”、间接性高冷等复杂交流场景时，能够准确理解意图，给出富有吸引力的回复。经过测试，在模拟的线上交流场景中，模型回复的满意度较初始版本提升 50%，有效帮助男性提升了情感交流质量。
  - **市场填补与用户价值创造：**填补了市场在精准化异性情感交流辅助模型领域的空白，为 18 - 35 岁面临线上交流难题的男性用户提供了针对性解决方案。
  - **技术积累与创新：**项目过程中积累了丰富的自然语言处理技术实践经验，尤其是在数据处理、模型微调、模型部署和应用开发方面。探索出一套结合多种技术的数据处理和模型优化流程，为后续相关项目的开展提供了宝贵的技术参考和创新思路，推动了情感交流辅助技术的发展。

2024.6-2024.8　　深圳市好邻养老科技有限公司　　VLM 老年认知唤醒项目

- **项目背景 ：**

  随着老龄化社会的加剧，老年痴呆等认知障碍问题日益凸显。据统计，我国 65 岁以上老年人中，约有 6% 患有老年痴呆症。认知障碍不仅影响老年人的生活质量，也给家庭和社会带来沉重负担。目前针对老年认知障碍的干预手段有限，传统的认知训练方式较为单一，缺乏趣味性和针对性。而大量存在的老照片，蕴含着丰富的记忆和情感信息，但仅靠人工引导回忆效率较低。VLM 模型结合图片数据，能够识别照片中的场景、人物等元素，通过生成与之相关的故事、问题，引导老年人回忆，刺激大脑活动，从而达到认知唤醒和训练的目的。
- **项目技术 ：**Deepseek-Vl-7b-Chat,Ms-Swift,Xinference
- **项目实施：**
  - **数据预处理：**对照片中的文字、场景、人物等关键信息进行提取，并将其转化为文本数据。针对 Deepseek-Vl-7b-Chat 模型进行微调前的数据预处理，将提取的文本数据整理成适合模型训练的格式，如将照片信息与对应的路径进行关联，为后续模型训练做准备。
  - **模型微调与优化：**以 Deepseek-Vl-7b-Chat 为基础模型，运用 Ms-Swift 微调框架，结合收集到的照片数据和对应的文本信息，对模型进行微调。在微调过程中，重点优化模型对照片场景、人物等元素的识别和理解能力，以及生成与之相关故事、问题的能力。通过设置不同的微调参数，如学习率、训练轮数等，进行多次试验，对比模型在生成故事、问题的逻辑性、趣味性和针对性方面的表现，选取最优的微调参数组合。利用开发的测试数据集，对微调后的模型进行性能评估，分析模型生成内容与照片实际信息的匹配度、对老年人认知唤醒的有效性等指标。
- **项目成果：**
  - **减轻家庭和社会负担：**该项目为家庭和养老机构提供了一种高效、便捷的认知训练工具，减少了对专业认知训练人员的依赖，降低了认知训练的成本。

- **创新认知训练模式**:突破了传统单一的认知训练方式,开创了基于老照片和 VLM 模型的新型认知训练模式。这种模式将回忆与认知训练相结合,具有趣味性和针对性,为老年认知障碍干预领域提供了新的思路和方法,有望成为未来老年认知训练的主流方式之一。
- **数据与模型资源积累**:在项目实施过程中,积累了大量经过标注和整理的老照片数据以及微调后的模型资源。这些数据和模型可作为珍贵的资源,为后续的相关研究和应用开发提供有力支持。通过对数据的持续优化和模型的进一步改进,有望形成具有行业影响力的数据与模型库,推动整个行业的发展。

2024.1-2024.6　　深圳市好邻养老科技有限公司　　老年在线学习领航项目

- **项目背景:**

  老龄化社会的到来,老年人对精神文化生活的追求日益增长,许多老人渴望学习新知识、新技能以丰富退休生活。然而,线上学习资源海量且分散,老年群体难以精准筛选出适合自己知识水平和兴趣的内容。同时,学习过程中遇到的问题也无法及时得到解答。该项目旨在利用 Agent 技术,帮助老年人便捷获取学习资源,提升学习体验。
- **项目技术:** Langchain,Langgraph,Tavily 搜索,GPT-4 API,Chromadb
- **项目实施:**
  - **Agent 核心逻辑构建**:基于 Langchain 框架搭建 Agent 的核心逻辑。定义 Agent 的决策机制,让其能够根据用户输入判断是否需要调用工具。例如,当用户输入“我想学习书法”,Agent 判定需要调用工具;若用户输入 “今天天气不错”,则作为闲聊场景,无需调用工具。
  - **Tavily 搜索工具集成**:将 Tavily 搜索 API 集成到 Agent 中,确保 Agent 可以在需要时调用该工具进行信息搜索。设置好搜索参数和调用规则,保证搜索结果的相关性和准确性。
- **项目成果:**
  - **精准的学习资源筛选:**当老年人提出学习需求时,Agent 能够准确判断并调用 Tavily 搜索工具,从海量的线上学习资源中筛选出适合他们知识水平和兴趣的内容。例如,一位想要学习绘画的老年人,在输入需求后,Agent 快速为其推荐了简单易学的绘画入门教程、适合初学者的绘画工具介绍等内容。
  - **良好的闲聊交互体验**:在闲聊场景中,Agent 能够与老年人进行自然流畅的对话,给予温暖、积极的回应。例如,当老年人分享自己的生活琐事时,Agent 会耐心倾听并表达理解和关心,让老年人感受到情感上的陪伴。通过良好的闲聊交互,增强了老年人与系统的亲近感和信任度,提高了他们使用系统的积极性。
  - **提升老年人学习积极性:**系统的便捷性和精准性帮助老年人更轻松地获取到适合自己的学习资源,解决了他们在学习过程中的困扰。许多老年人表示,使用该系统后,他们对学习新知识、新技能的兴趣明显提高,愿意投入更多的时间和精力进行学习。

2023.7-2023.12　　深圳市好邻养老科技有限公司　　老年健康知识咨询项目

- **项目背景:**

  随着老龄化加剧,老年人健康问题频发,可他们获取准确健康知识的途径有限,网络信息繁杂难辨,难以满足个性化需求。而我们的 RAG 系统凭借精准检索与智能生成,为老人提供贴合其需求的可靠知识,成为他们健康管理的得力助手。
- **项目技术:** Llamaindex,Chromadb,GPT-4 API
- **项目实施:**
  - **数据准备与预处理:**
    - **数据收集**:为了满足老年人对健康知识的需求,收集涵盖常见老年疾病(如高血压、糖尿病、冠心病等)的预防、治疗、护理知识,养生保健方法,合理饮食搭配,运动建议等方面的专业医学书籍、学术论文、权威健康科普文章等资料。将这些资料整理存放在指定的文件夹中,方便后续使用 SimpleDirectoryReader 函数进行读取。

- **文件读取与分块**：使用 SimpleDirectoryReader 函数从指定文件夹中读取所有文档数据。设置合适的分块参数，如 Settings.chunk_size = 512 和 Settings.chunk_overlap = 25，将文档内容分割成适当大小的文本块，以便后续进行向量化处理。

- **知识索引与模型集成搭建:**
  - **向量库与索引构建**：首先初始化 ChromaDB 客户端，连接到指定服务，设置集合名称为 “rag”。接着，运用 HuggingFaceEmbedding 模型，对分割后的文本块进行向量化处理，将文本转化为向量以便高效存储和检索。之后检查向量存储情况，若不存在则创建新的存储，把向量化后的文本块存入 ChromaDB 向量库，使用 VectorStoreIndex.from_documents 函数构建索引，并将其存储在本地，方便后续快速加载。
  - **大语言模型集成**：从环境变量获取 API Key 后，初始化 OpenAILike 大语言模型，设置 API 地址、模型名称、温度、最大 token 数等参数，保证模型正常运行并能生成准确回复。然后将初始化好的大语言模型和嵌入模型设置到 Settings 中，供后续查询和对话使用。
- **对话交互与信息持久化:**
  - **对话引擎构建与交互**：使用 index.as_query_engine() 构建查询引擎，用于处理用户单次查询请求，能依据用户问题在向量库检索信息并生成回复。利用 index.as_chat_engine 函数构建聊天引擎，采用 condense_question 模式，设置匹配相似度前 5 的知识库片段，开启详细日志打印，以方便调试优化。通过无限循环实现多轮对话，持续接收用户问题并调用聊天引擎回复。
  - **对话信息持久化存储**：利用 ChromaDB 来持久化存储对话信息。可以为对话记录单独创建一个 ChromaDB 集合，例如命名为 "conversation_records"。每次对话结束后，将用户提问、系统回复、对话时间等信息组合成一个文本段落，然后使用嵌入模型（如之前的 HuggingFaceEmbedding）对该段落进行向量化。将生成的向量和对应的元数据(如用户 ID、对话 ID 等)存入 "conversation_records" 集合中。这样，后续可以基于 ChromaDB 的检索功能对对话历史进行分析统计，例如根据用户提问的关键词检索相关对话、按照时间范围筛选对话等，进而优化系统性能和服务质量。

◆ **项目成果:**

- **精准的健康知识查询服务**:系统能够根据老年人的提问，从存储的大量健康知识文档中精准检索相关信息，并生成准确、易懂的回复。例如，当老年人询问 “高血压患者日常饮食需要注意什么” 时，系统可以快速从向量库中找到相关内容，详细介绍高血压患者应遵循的饮食原则，如低盐、低脂、高纤维等，并列举适合和不适合的食物种类。经过测试，系统对常见健康问题的查询准确率达到了 85% 以上，大大提高了老年人获取准确健康知识的效率。
- **良好的多轮对话体验**：采用 CondenseQuestionChatEngine 构建的聊天引擎实现了自然流畅的多轮对话功能。老年人可以在对话过程中逐步深入探讨健康问题，系统能够根据上下文理解用户的意图，并给出连贯的回复。例如，在询问高血压饮食注意事项后，老年人可以进一步询问 “有没有适合高血压患者的食谱”，系统能够结合之前的对话内容，提供相关的食谱建议。用户反馈显示，系统的多轮对话交互体验良好，能够满足他们在获取健康知识过程中的交流需求。
- **长期记忆与服务优化**：通过对话信息的持久化存储，系统形成了长期记忆，能够对用户的提问历史进行分析和统计。例如，分析用户关注的健康问题类型、提问频率等，以便更好地了解用户需求，为用户提供个性化的健康知识推荐和服务。基于长期记忆数据，系统可以不断优化向量库的索引结构和检索算法，提高查询效率和回复质量，进一步提升服务的精准性和实用性。