

个人简历

段佳兵

duanjiabingcom@163.com | 19313069493 | 2002/02/19 | 后端开发工程师



教育背景

华南理工大学

计算机科学与工程学院 · 计算机科学与技术 · 硕士

2024.09 - 2027.06

广州

- 核心课程：高级计算机网络 (89)、高级数据库系统 (86)、机器学习算法 (86)、高性能计算技术 (83)
- 研究方向：大模型分布式训练场景下的 RDMA 网络流量负载均衡

中南大学

机电工程学院 · 微电子科学与工程 · 学士

2019.09 - 2023.06

长沙

- 核心课程：理论力学 (100)、线性代数 (95)、C++ 程序设计及应用 (94)、微机原理及应用 (88)
- 在校期间成绩优异 GPA: 85.12/100, CET-4/6, 并获得多个比赛奖项

专业技能

C++ 后端 高并发网络编程, Redis, MySQL
AI 基础设施 llama.cpp 推理部署, RAG, Agent 架构
语言与工具 C++ (主力), TypeScript, CMake, GDB, Git, Linux

实习经历

腾讯科技有限公司

后台开发实习生 · IEG · 光子工作室群 · 《掌上无畏契约》服务端组

2026.07 - 2026.10

深圳

- 昵称生成平台：面向游戏策划的批量命名需求，基于 React、FastAPI、SQLite 开发集昵称分析、生成与评估于一体的 Web 平台，通过后台 Worker 执行任务并持久化状态，接入 Chat Completions API 与 Batch API，支持在线与批量生成。
- 玩家风格分析：针对短昵称字面相似性掩盖风格差异的问题，设计 LLM 风格扩写 + Embedding + KMeans 流程，在同一批 1 万条昵称上，将轮廓系数由 0.0171 提升至 0.1264。
- 命名素材提取：基于 10 万条真实玩家昵称，结合 N-gram 候选挖掘、人工筛选与 Jieba 自定义词典，保留游戏术语、梗及命名结构片段，2/3/4 字片段 Recall@100 达到 0.94/0.86/0.91。
- 生成策略与评估：将聚类结果提炼为表达机制卡与风格桶，结合词频、人设及 Few-shot 构建生成指引；建立用词分布、语义相似度、LLM-as-a-Judge 三维评估体系，通过对照与消融实验权衡生成质量和 Token 开销。
- 生成重复优化：针对固定 Prompt 引起的重复，结合批量生成、动态 Few-shot 与排除列表，将原始唯一率由 28.3% 提升至 89.6%。

项目经历

基于 C++ 的实验室私有化大模型问答平台

2025.11.1 - 2026.5.20

开发面向实验室内部的全私有化 AI 问答平台，支持检索增强生成，满足不同方向同学并发使用，同时做到严格的权限隔离和 Token 计费。

技术栈：C++ | Reactor | RAG | llama.cpp | MySQL | Redis

- 网络层与协议实现：利用 epoll 实现一主多从 Reactor 多线程模型。每个工作线程独立绑定 epoll 实例，连接轮询分发后绑定同一线程，天然无锁。利用主从状态机解析 HTTP/1.1 协议，支持分块传输编码解码、跨域请求处理。
- 大模型推理引擎：集成 llama.cpp 推理框架，启用 CUDA 加速编译。设计会话级 KV 缓存淘汰算法，按 LRU 在会话间自动回收显存。实现系统提示前缀共享机制，多个会话共用相同系统提示时复用缓存前缀。
- RAG 检索增强生成：实现完整 RAG 管线，利用 FAISS 实现向量检索，BM25 实现关键词检索，利用 RRF 进行倒序重排，结合元数据增强实现过滤。最终上下文精确率 85%+。
- 实验室安全与权限控制：JWT 绑定多租户身份，颁发 24 小时有效令牌，内含用户标识、租户标识及角色信息。所有数据库查询自动注入租户过滤条件，杜绝跨租户数据泄露。

基于 Nanoclaw 架构的多智能体学术科研助手

2026.4.1 - 2026.6.10

基于 Nanoclaw 框架的 Multi-Agent 学术助手，采用 Node.js 宿主 + Docker 沙盒架构，支持文献自动检索、深度阅读、对比归纳与定时推送，以及科研思路深度探讨。

技术栈：Node.js | TypeScript | Docker | Plan-and-Solve | ReAct

- Multi-Agent 协同架构：设计 Plan-and-Solve 范式的主 Agent 负责复杂指令拆解与决策路由，多个 ReAct 范式子 Agent 负责垂类细分任务。利用 ToolRegistry 工具挂载机制实现主从 Agent 间的调度与上下文数据传递。
- Docker 安全沙盒与 IPC 通信：宿主调度器接收指令，动态派发给隔离的 Docker Python 沙盒执行 Agent 任务。利用轻量级文件 IPC 通道安全回传 Agent 思考与结果，通过文件监控识别并返回对话源，Docker 隔离有效防范 LLM 恶意代码越权。
- 飞书交互枢纽与多会话控制：接入飞书对话源，在宿主端基于 ChatID 实现组级路由，结合动态分配虚拟挂载目录至对应 Group，实现多会话安全隔离与上下文管理。

荣誉奖项

第二十届大学生机器人大赛 ROBOMASTER 机甲大师高校单项赛“工程采矿”项目全国一等奖

2021

第二十届大学生机器人大赛 ROBOMASTER 机甲大师超级对抗赛（南部赛区）一等奖

2021

省级大创项目“基于 STM32 的仓库管理机器人控制及相关研究”负责人，获优秀结题，优秀案例等

2020

第十一届 APMCM 亚太地区大学生数学建模竞赛二等奖

2022