张皓泉 AI应用全栈开发（前端侧重）

意向城市：杭州/北京 | 电话: 13388005839 | 邮箱: 1519815799@qq.com
作品集: link2ai.online | GitHub: @linkzhangithub | 技术博客: 语雀博客

简 个人总结

- 日常用 Cursor、Trae、DeepSeek 辅助开发，在 Trae 上实践过6A工作流（需求对齐→架构设计→任务拆分→审批→执行→验收），习惯用标准化流程让 AI 按节奏交付。
- 独立完成3个完整上线的 AI 项目（Vue3 + Node.js），通过腾讯云 EdgeOne Pages 部署展示。
- 写代码的同时重视用户体验：流式消息滚动时不卡、断网不白屏、图片上传时先压缩减少用户等待时间。
- 习惯先把复杂问题拆解清楚，理清依赖关系和优先级，再按步骤执行。

项 项目经历

RAG 知识库问答系统 独立开发

2026.04 - 至今

项目概述

基于 RAG 技术实现基于文档内容的精准问答，支持在线演示（预置文档）与本地部署（自定义文档）双模式。

核心功能

支持 PDF/TXT/MD/DOCX 多格式文档上传；查询扩展与同义词映射增强召回；三路混合检索（向量相似度+关键词覆盖率+人名加权）+ 两阶段排序优化相关性；SSE 流式输出 + 心跳保活。

Vue3 | Vue Router | Node.js | Express | 向量相似度计算 | LLM API集成 | SSE流式传输

- **多格式文档解析兼容性：**PDF 和 DOCX 解析时遇到中文乱码、表格丢失等问题。通过组合使用 pdf-parse 和 mammoth.js，针对不同格式选择最优解析器。上传环节通过 iconv-lite 实现文件名编码自动检测，支持 UTF-8、GBK、GB2312 等多种编码转换，解决了中文文件名乱码问题。
- **检索召回率低，答非所问：**向量检索对专业术语召回差、停用词过滤过度。采用查询长度自适应混合检索（长查询≥8字：向量60%+关键词30%+人名10%；短查询< 8字：向量50%+关键词40%+人名10%），并通过 expandQuery 同义词映射表扩展查询意图（如"rag"自动扩展为"检索增强生成 RAG 知识库问答"），优化停用词表。两阶段排序：先混合检索取候选集，再用关键词覆盖度 Rerank（混合分数80%+覆盖度20%），显著提升相关性。
- **大模型幻觉问题：**LLM 可能编造文档外内容。通过 temperature=0.1 + top_p=0.5 压缩采样随机性，并在 Prompt 中设置多层约束（禁止编造、禁止外泄知识、兜底话术"根据现有资料无法回答"），同时在文件名拼接 Prompt 前通过 sanitizeFileName 过滤特殊字符，防止格式注入干扰模型输出。
- **API 调用不稳定：** 向量化和 LLM 推理依赖外部 API，网络波动导致请求失败。实现指数退避重试机制（最大重试 3 次，基础延迟 1 秒），并对错误分类处理（4xx 客户端错误立即返回、429 限流/5xx/网络错误自动重试）。同时接入 express-rate-limit 做接口限流（每分钟 10 次），防止恶意调用。
- **SSE 长连接断开：**RAG 问答采用 SSE 流式返回答案，长时间无数据传输导致连接被代理服务器断开。增加心跳保活机制（每 15 秒发送一次心跳包），保持连接稳定。

项目经历（续）

AI 写作助手　独立开发　2026.03 - 至今

项目概述

解决长文写作大纲构思难、扩写耗时问题，实现从大纲生成、分节扩写到 AI 质检的完整闭环。

核心功能

输入主题自动生成结构化大纲；分节内容一键生成，支持润色/扩写/缩写；上下文感知生成确保全文一致；AI 质检（5维度评分+改进建议）；打字机效果展示，支持随时中断。

Vue3　Pinia　Vite　Tailwind CSS　Node.js　Express　LLM API集成

- **大模型响应延迟导致用户体验差**：LLM 生成内容时响应较慢，用户等待焦虑。正文生成引入 SSE 流式传输，将生成过程分块返回；前端配合打字机动画逐字展示，让用户可提前阅读。同时基于 AbortController 实现请求中断控制，用户可随时停止生成。
- **长文本生成时前端渲染卡顿**：扩写长内容时，频繁的 DOM 更新导致页面滚动和交互卡顿。采用分章节逐个生成策略，每次只渲染当前章节内容，配合 Vue 的 shallowRef 减少不必要的响应式追踪，保持页面交互流畅。
- **大纲与正文内容结构不一致**：用户修改大纲后重新生成正文，容易出现前后内容矛盾。采用大纲驱动的生成策略，每次生成都基于当前最新大纲结构，确保正文内容与大纲保持一致。
- **AI 质检返回 JSON 格式不稳定**：LLM 输出的评分 JSON 可能包含 Markdown 标记或格式异常。实现多重清洗（正则去除代码块标记 + JSON 范围精确截取）、兼容新旧字段格式（suggestions 字符串/对象数组统一转换）、分数边界保护（Math.min/Math.max 钳位至 0-20），解析完全失败时返回默认评语而非白屏，确保前端始终可用。

AI 对话机器人　独立开发　2026.01 - 至今

项目概述

解决通用对话缺乏场景适配问题，通过预设角色和语音交互提升对话体验。

核心功能

预设 4 种角色适配不同场景；SSE 流式输出减少等待感；语音输入转文字；图片理解功能；对话历史持久化与多轮上下文。

Vue3　Vue Router　Pinia　SSE　Markdown渲染　代码高亮　语音识别SDK

- **角色切换时对话风格不连贯**：动态注入角色system prompt确保风格一致，基于Pinia + localStorage持久化多轮对话历史，支持切换角色不影响对应上下文。
- **对话内容渲染安全性**：采用 MarkdownIt（禁用 HTML）+ highlight.js 解析渲染，从源头禁止危险标签，在支持富文本展示的同时避免 XSS 漏洞。同时在 system prompt 中约束 AI 不返回 HTML/JavaScript，双管齐下。
- **语音识别稳定性**：集成讯飞语音识别 SDK，通过 onWillStatusChange 监听状态流转（idle→ing→end）确保 UI 麦克风图标与录音状态严格同步；配置缺失时前端弹窗提示，识别异常时自动调用 stopListening() 回收资源并重置状态，避免多次点击创建多个实例导致内存泄漏。
- **图片理解响应慢**：Canvas 压缩图片至 800px/500KB 再上传，前端限制 5MB，优化等待体验。

技 专业技能

前端开发： Vue3 · TypeScript · Vite · Pinia · Vue Router · Tailwind CSS · 响应式设计

AI 应用： LLM API集成 · RAG检索增强 · Prompt Engineering · 向量检索 · 多模态识别

后端开发： Node.js · Express · RESTful API · SSE流式输出 · 限流控制

工程化与部署： Git · 腾讯云EdgeOne · CI/CD基础

教 教育背景

天津职业技术师范大学 本科 软件工程 2019年 - 2021年 | 天津

- 全日制统招本科，工学学士学位

天津现代职业技术学院 专科 会计电算化 2016年 - 2019年 | 天津

证 证书荣誉

人社部 AIGC 应用工程师 - 2026年

工信部 程序设计师 - 2021年

"互联网+"大学生创新创业大赛校三等奖 - 2021年

校园各类奖学金 - 2017-2021年

工 工作经历

明源云（天津） 软件实施工程师 2021年 | 天津

- **业务流程配置：**驻场天津NCC新城市中心项目（客户：天津西青区中北镇政府），基于明源云空间平台完成租约审批、资产台账、收费规则等核心流程配置。针对系统与客户业务的差异，通过字段扩展、表单重构、规则编排等手段进行低代码个性化适配；主导分析并拉通产研推动迭代。
- **需求协调与管理：**依负责客户需求的全链路转化，在Jira中完成需求创建与状态跟踪，协同管理40+条需求池。面对需求增速高于研发产能，通过优先级排序与预期管理，确保核心诉求落地。
- **项目交付支持：**参与系统部署、流程验证及UAT测试。在客户发现前主动识别并修复8条重大bug，规避上线事故与投诉风险。协助系统上线验收，推动平台平稳过渡至运营阶段。

金融咨询创业项目 创业合伙人 2022年 - 2025年 | 天津

- **客户服务与资质审核：**负责客户信息核验与贷款申请资料的完整性审查，协助客户完成线上申请操作全流程，确保资料规范、合规、齐全，提升申请通过率。
- **团队协同与流程运营：**参与团队分工协作，涵盖短视频引流、客户邀约、信息核验等环节，打通前端获客到后端转化的服务链路，累计服务客户数百人。
- **进度追踪与风控闭环：**跟进贷款审批全流程，及时向客户同步审核结果，确保审批进度透明可控，实现从申请提交到放款结果的高效反馈闭环。