

李浩哲

HAOZHE
LI

概述

李浩哲是一位经验丰富的AI工程师，擅长大规模LLM（大语言模型）的实施和优化，专注于多智能体系统和 Retrieval-Augmented Generation (RAG) 技术的应用，以高效的产品开发和部署方式。同时，他精通基于Kubernetes的容器编排和高性能推理框架，如NVIDIA Triton等，拥有从架构设计到全栈解决方案部署的全栈开发能力。

实习经历

AI工程师

华为 | 2025.5-2025.8

- 加入云数据库创新Lab，深度参与多模态RAG As A Service概念验证项目，负责文档解析与语义嵌入的选型与开发。
- 主导基于VLM及视觉模型的文档解析流水线设计，召回率提升**10%**，解析性能提升**18%**。
- 研发语义与结构混合切块策略，结合元数据上下文，配合延迟切块使系统召回率提升**7%**。
- 参与混合嵌入/检索、Dynamic Batching及高性能推理框架的开发与集成。

AI工程师

RPGGO AI | 2024.5-2024.8

- 独立开发并优化了流式TTS和语音克隆系统，整合多家API实现高并发低延迟。
- SFT大模型辅助音乐生成过程，以匹配游戏主题和场景。

教育

计算机科学与音乐 理学士

伊利诺伊大学香槟分校
2022-2026

- 伊利诺伊数学实验室软件工程师：参与开发 ColorTaiko!项目，一款受代数拓扑研究启发的益智游戏
- CS124课程助教：课程辅导答疑
- 核心课程：数据库系统、机器学习、数据科学、网页设计
- 院长清单

项目

Omni: 深度研究Agent

项目链接: omniknows.xyz

基于Hierarchical Supervisor Multi-Agent框架的 ReAct AI复合智能体，具备实时联网、深度推理与自主编程调试能力 缓存网页至自托管Milvus向量数据库，在非时效性查询场景下相较直接联网搜索提升约48%速度 后端使用Langgraph构建，全部容器化实现分布式部署，并通过Cloudflare CDN进行边缘计算，配合自动预热与保活进一步

Chat Syllabus

项目链接: chatsyllabus.xyz

领导四人团队，探索大模型和 RAG 在教育领域的应用，构建智能课程辅助系统，实时解析并回答课程相关问题。设计预测模型，自动评估课程难度和学习负担，助力学生制定学习计划。

ZeroAI: AI生成内容检测

项目链接: zeroai.haozheli.com

通过计算文本困惑度并结合随机森林模型，可精准预测文本是否为 AI 生成，F1-Score **97%**。将 GPT-2 Distill模型量化后转为ONNX Runtime，降低冷启动时间约**70%**，CPU推理性能提升为**22%**。

技能

编程语言

Python、JavaScript、C++、Java、Golang、SQL

AI / ML

增强检索生成、AI Agent系统、提示工程、大模型应用开发、TTS / ASR、PyTorch

工程Infra

Kubernetes、NVIDIA Triton、Docker、FastAPI

网页设计开发

React、NextJS、Figma

其它 / 软技能

视频剪辑、混音、熟练使用ChatGPT

联系

MAIL@HAOZHELI.COM

WWW.HAOZHE.LI

LINKEDIN.COM/IN/HAOZHELI

+86 156 0123 3689