# 王俊豪

1305825523@qq.com (+86) 13666418143 A13666418143

## 教育背景

**东北林业大学 211 院校 双一流** 2023 – 2027

本科在读 人工智能

## 专业技能

- 熟悉 **RAG 全链路**，掌握文档解析、Chunk 切分、混合检索、Rerank、上下文拼接、引用核查与效果评估
- 熟悉 **Agent 应用开发**，了解 Query Loop、Tool Calling、任务规划、上下文管理、权限控制与执行记忆机制
- 熟悉 **Claude Code**、**OpenClaw**、**Hermes** 等 Agent 工具，能够结合项目场景完成代码理解、任务拆解与开发辅助
- 熟悉 **Python** 与大模型应用工具链，具备 LangChain、Chroma、BM25、RAGAS、OCR/VLM 等实践经验
- 熟悉 **PyTorch**，了解计算机视觉模型训练、微调与模块优化流程

## 项目经历

**MiniCode** **AI 应用开发** 2025.12 – 2026.04

**项目简介：**参考 Claude Code 类 Coding Agent 的执行范式，设计并实现面向小任务执行的 AI Coding Agent MVP，支持项目结构分析、报错分析、小功能实现计划与局部 Patch 建议，重点验证 Skill 路由、上下文压缩、权限审查与执行记忆机制。

**技术栈：** Agent、Tool Calling、Skill Router、Context Management、Memory System、Python

**技术亮点：**

- **Query Loop 执行闭环：**构建“任务识别—Skill 路由—工具调用—上下文构建—模型生成—权限审查—结果输出”的任务执行链路，实现 Agent 单轮任务的状态管理与过程追踪
- **Skill 能力体系：**将 explain_project、fix_error、small_feature_plan、patch_suggestion 等高频任务抽象为 Skill，通过任务意图、关键词与元信息完成候选召回，降低模型直接选工具的不稳定性
- **上下文压缩：**将大文件和工具结果压缩为“文件路径、功能摘要、关键片段、占位 ID”的结构化上下文，减少无关信息干扰，提升长任务上下文稳定性
- **权限与安全审查：**封装 list_files、read_file、search_code 等工具，并加入路径校验与风险过滤，默认禁止读取.env、私钥、token 等敏感文件
- **执行记忆沉淀：**将任务类型、错误模式、解决经验、相关文件与执行结果保存为结构化记忆，用于后续同类任务的经验复用

**面向个人的多模态 RAG 知识库问答系统** **AI 应用开发** 2026.01 – 2026.05

**项目简介：**面向个人知识管理场景设计并实现多模态 RAG 问答系统，支持 PDF、Markdown、图片等知识源，围绕离线索引、在线检索、效果评估、增量索引与缓存优化完成全链路工程实践。

**技术栈：** Python、LangChain、Chroma、BM25、RAGAS、OCR、VLM、Rerank

**技术亮点：**

- **多模态离线索引：**优化 PDF、Markdown 解析流程，结合 OCR 与 VLM 提取图片文本和语义信息，并对 Chunk 进行清洗、去噪、元数据绑定与索引构建
- **混合检索与精排：**构建 BM25 + 稠密向量的混合检索链路，结合 Rerank 对候选 Chunk 进行相关性精排，提升关键词问题和语义问题的召回效果
- **在线问答优化：**设计查询改写、问题路由、多路召回、上下文拼接与引用核查机制，并在低置信度场景下触发二次检索，降低无依据回答
- **评估与调优：**基于 RAGAS、MRR、Hit@K 等指标搭建评估流程，对 Chunk 大小、Top-K、召回策略和 Rerank 参数进行对比优化
- **增量索引与缓存：**基于文档 Hash 实现增量索引与热更新，支持文档新增、修改、删除后的快速同步，并通过分层缓存复用解析结果、索引结构与高频问答结果