



肖俊鹏

17750091909 | 2568111960@qq.com

男 | 年龄：24岁



北京航空航天大学
BEIHANG UNIVERSITY

教育背景

北京航空航天大学 (985) 硕士 (保送) 机械工程 (计算机图形学方向) 2024.09-至今

方向：数字化设计与图形学。

获奖情况：学业二等奖学金、学业三等奖学金

中国农业大学 (985) 本科 计算机科学与技术 (理科试验班) 2020.09-2024.06

毕业设计：《基于unity的数字博物馆的设计与开发》

获奖情况：美国大学生数学建模竞赛M奖

实习经历

引擎研发-AI Infra实习生|Unity中国公司 2026.6-2026.9

简介：设计并实现 Unity 动画自然语言检索工具，融合动作采样、Caption RAG、TMR 动作向量与结构化重排。

1、动画数据与特征：实现 FBX/AnimationClip 扫描、统一骨架管理、Clip/Segment 切分及逐帧采样，构建**47996条**动画资源库和1,000条实验集，生成8,000个时序特征区间

2、TMR 动作向量：打通 **Unity 22关节位置序列** → **Guo 263维特征** → **TMR 256维 Motion Embedding** 管线，实现自然语言对 Clip 及多尺度 Segment 的直接检索

3、混合检索：实现 Caption RAG、TMR、TMR + Structured Re-rank 三种模式，融合文本相似度与速度、方向、姿态等动作特征；**原型集检索 Top1 达 87.1%**

4、工具与评测：开发 Unity Editor 检索窗口，支持评分解释、Segment 预览和报告导出；建立 Top1/Top3、方向准确率及查询耗时评测，并实现开发集/盲测集 零语义泄漏划分。

大模型推理框架实习生|原粒半导体公司 2026.3-2026.6

简介：基于 llama.cpp 的自研推理后端适配

1、为 llama.cpp Server接入自研芯片推理后端，将ggml graph执行路径扩展为编译模型Runtime，保留Batch、Slot、Sampler与流式输出逻辑；完成 2个主流模型、2种量化格式适配，贪心解码下端到端 **Token一致率达 99.3%**

2、适配Prefill/Decode、Sequence/UBatch拆分、设备侧KV Cache、Buffer映射及Logits重排，支持 Batch 1~8、**上下文长度最高8K**；相比初版后端，**推理吞吐提升约 22%、首Token延迟降低约 18%**

3、通过Adapter兼容llama.cpp Memory接口，打通文本生成、Embedding和Rerank三类任务；完成 **28项回归测试及24小时稳定性测试**，后端推理结果可无缝接入Token采样与Server流式响应流程

大模型应用研发实习生|上海天跃公司 2025.6-2025.12

简介：负责安防领域智能助手 App 的核心研发

1、对话编排与意图识别优化：基于 LangChain设计并实现了针对安防垂直领域的意图识别模型。引入 **Few-shot Prompting 与动态上下文注入**，将复杂多轮对话中的场景**识别准确率从78%提升至 94%**。针对用户模糊表达，设计了反向提问机制，有效降低了对话的中断率。

2、RAG 检索增强与知识库优化：针对安防知识库中存在的误报、漏报等长尾问题，优化了向量检索与重排序策略。通过对安防设备手册进行精细化分块 (Chunking策略优化)，使知识库问答的**召回率 (Recall) 提升了22%，幻觉率降低了 15%**。

3、高性能工程落地与性能监控：利用 FastAPI 封装大模型接口，引入 SSE流式响应技术，将用户感知的首字响应延迟 (Time To First Token) 从**3s+压缩至0.8s以内**。

项目经历

面向航空发动机装配领域的多源知识图谱 RAG 框架

MDPI在投 第一作者

2026.1-2026.6

简介：基于航空发动机装配领域存在 BOM 表、维修手册、CAD模型三源数据异构、装配知识分散的现状，设计并独立实现了一套多模态知识图谱增强问答系统。

1、构建三源知识图谱：开发 LangGraph 四阶段提取管道，从 DOCX 格式 BOM 表（结构化解析）、维修手册 PDF/Markdown (LLM实体关系抽取, 460 段落)、STEP 格式 CAD 文件 (Open CASCADE 装配树解析) 中提取知识写入 Neo4j，**达到 7333 节点 / 8752条边**，覆盖9 类工程关系

2、实体对齐与质量提升：设计四级 BOM 对齐策略（零件号提取 → 图号噪声清洗 → 双向子串 → Token交集匹配），**相比纯模糊匹配误判率下降约 12%，Part 类三元组对齐率达 34%**，为装配问答提供可靠的结构化锚点

3、混合检索问答架构：实现文本向量 (bge-m3) + BM25 + Cross-Encoder 重排序的**混合检索**路径，与 Neo4j 图谱查询 (Kahn拓扑排序工序链 + 递归 BOM 子树) 双路融合，通过 **SSE 流式生成** 带来源溯源的装配方案回答；在自建 200 题分级 QA数据集上评测，**混合检索 Recall@5 达 100%、MRR 达 0.99**，覆盖结构 / 工序 / 规格 / 几何四类问题

4、评测体系：构建 200 题分级 QA 数据集，覆盖 BOM层级、工序推理、规格查询、跨源对齐四类能力，采用 LLM-Judge + RAGAS 双路评估

项目链接：<https://github.com/xjp-buaa2025/rag-demo>

校园经历

- 学生会文艺部部长(2021.9-2022.9) •交互式编程技术课程助教(2024.9-2025.1)
- 工业软件基础课程助教(2025.3-至今)

专业技能

编程语言：Python、C#、C++，熟悉 FastAPI、LangChain/LangGraph、Unity Editor 开发
AI与检索：RAG、Embedding、BM25/RRF、HNSW、Cross-Encoder、知识图谱、Agent编排
大模型工程：llama.cpp、vLLM、模型推理服务、KV Cache、Prefill/Decode、流式生成
工程工具：Linux、Docker、Git、Neo4j、WSL2