孔维轩 (Weixuan Kong)

Jackkong29@Gmail.com 18796611232 github.com/Jackela Portfolio

教育经历 (Education)

悉尼大学 (The University of Sydney) *2024 年 2 月 – 2025 年 12 月 (预计)*
计算机科学硕士 *(Master of Computer Science)* *WAM: 77.4 / 100*

阿尔戈马大学 (Algoma University) *2020 年 9 月 – 2023 年 6 月*
计算机科学学士 *(Bachelor of Computer Science)* *WAM: 88.8 / 100*

荣誉与认证 (Honors & Awards)

- **荣誉学位 (Cum Laude)** - 阿尔戈马大学, 2023: 因卓越的学术表现（专业排名前 10%）被授予此荣誉。

核心技能 (Core Competencies)

架构与云原生	**领域驱动设计 (DDD)**, 微服务, 分布式系统, 事件驱动架构, 无服务器 (Serverless), **Docker**, **Kubernetes**, **AWS**, 基础设施即代码 (IaC)
AI 核心技术	**检索增强生成 (RAG)**, 大语言模型应用 (LLM), 多智能体系统, 向量数据库, A/B 测试, Haystack, Langchain4j
语言与工程实践	**Java**, **Python**, JS/TS (React/Next.js), **Spring Boot**, FastAPI, **TDD**, CI/CD, SQL/NoSQL, 敏捷开发 (Agile)

实习经历 (Internship Experience)

北京愿景明控集团 (Beijing Vision and Control Group) *2023 年 6 月 – 2023 年 8 月*
人工智能团队实习生 *(AI Team Intern)*

- **项目背景:** 团队核心业务使用 Java 技术栈，而前沿的 AI 研究则依赖 Python 生态，两者间的技术壁垒导致新 AI 功能的集成与验证周期长、效率低下。
- **解决方案与实现:** 独立设计并工程化实现了一款 Java-Python **服务桥接工具**。该工具允许不熟悉 Python 的 Java 工程师通过简单的 Java 方法调用，无缝地、安全地与 Python AI 服务进行高性能通信。
- **核心贡献:** 将 AI 模块的**集成时间从预估的数天缩短至数小时**，直接推动了 AI 音箱产品原型**提前 1 周**完成并进入验证阶段，极大地加速了团队对 AI 方案可行性的评估。

核心项目经历 (Flagship Projects)

StoryForge AI *2025 年 7 月 – 至今*
系统架构师 & AI/DevOps 工程师 ——一个基于 *Kubernetes* 部署的、高可用的云原生多智能体 *AI* 叙事平台

- **项目背景:** 旨在解决传统 AI 叙事缺乏连贯性与“意外感”的核心痛点，通过模拟一个包含多个自主 AI 代理（导演、角色、史官）互动的“世界”，以动态生成和记录“涌现式”的复杂叙事。
- **架构设计:** 主导设计了一个云原生、分布式多智能体系统。为实现企业级部署和高可用性，使用**基础设施即代码 (Terraform)** 来定义和自动化部署所有云资源，并使用 **Kubernetes (k8s)** 对包含 AI 服务、API 网关、监控栈在内的整个系统进行容器编排。

- **技术挑战与解决方案:** 为解决 LLM 的长期记忆与上下文窗口限制问题，设计并实现了一套包含工作记忆、情节记忆和语义记忆的**分层记忆模块**。通过动态上下文构建器，在每次 AI 决策前从不同记忆层中检索最相关信息，最终将平均每次 AI 决策的 **Token 消耗降低了 75%**。
- **项目成果:** 通过 Terraform 和 Kubernetes，将完整的生产环境**部署时间从数天缩短至 30 分钟**。在混沌工程测试中，系统展现了高可用性，Kubernetes 能在 1 分钟内自动恢复被模拟杀掉的服务。项目不仅是一个技术展示，更是一套可复用的**企业级云原生部署蓝图**。

AI Enhanced PDF Scholar

2025 年 6 月 – 2025 年 8 月

架构师 & 全栈开发者 ——*一个基于清洁架构 (Clean Architecture) 与 RAG 的生产级智能文献分析平台*

- **项目背景:** 旨在解决学术研究者在处理海量 PDF 文献时工具链碎片化、工作流中断的核心痛点，将静态的阅读体验转变为动态的、智能化的知识探索过程。
- **架构设计:** 主导了项目从单体桌面应用到现代 Web 服务的架构转型。新架构采用前后端分离模式（React + FastAPI），并严格遵循**清洁架构 (Clean Architecture)** 和 SOLID 原则，实现了业务逻辑与基础设施的完全解耦，极大地提升了系统的**可测试性**与**可扩展性**。
- **技术挑战与解决方案:** 为解决 RAG 查询的计算与内存密集问题，设计并实现了一个包含**优先级队列**和**并发工作线程池**的异步任务管理器，并开发了能在任务执行前检查系统负载、执行中监控内存并主动触发 GC 的**内存安全处理器**，确保了系统在高负载下的稳定性。
- **项目成果:** 成功将核心 API 的 P95 响应时间优化至 **210ms**，RAG 平均查询时间稳定在 **4.2 秒**。通过智能缓存设计，将对外部 LLM API 的重复调用**减少了约 40%**。项目不仅交付了一个产品，更沉淀了一套**可移植的监控与告警框架**和企业级项目文档体系。

CATAMS (CAPSTONE 毕业设计)

2025 年 8 月 – 至今

全栈开发者 & *DevOps* 工程师 ——*一个基于领域驱动设计 (DDD) 和 TDD 的全栈企业级工时管理系统*

- **项目背景:** 旨在解决高校普遍存在的、依赖手动流程的临时工时管理难题，该流程效率低下、易于出错且难以审计。
- **架构设计:** 后端采用 **Java 21 + Spring Boot 3**，并严格遵循**领域驱动设计 (DDD)** 的思想。将复杂的审批逻辑（涉及至少 7 个状态）封装在核心领域对象 'ApprovalStateMachine'（**状态机模式**）中，确保了业务规则的一致性和健壮性。
- **技术挑战与解决方案:** 为应对异构技术栈（Java/Node.js）的测试挑战，编写了一套 **Node.js 脚本作为所有测试的统一编排层**。为保证集成测试的可靠性，引入 **Testcontainers** 库，在运行测试时会自动在 Docker 中启动一个临时的 PostgreSQL 实例，保证了每次测试都在纯净、隔离的环境中运行。
- **项目成果:** 交付了一个业务逻辑零缺陷的健壮系统，并通过全面的自动化测试（单元、集成、契约、E2E）为每次部署提供了极大信心。项目构建了一套**可移植的全栈自动化测试框架**，并沉淀了**企业级现代 Java 应用**的开发最佳实践。

其他项目经历 (Other Selected Projects)

AI Recruitment Clerk

2025 年 6 月 – 至今

解决方案架构师 & 后端技术负责人

- 主导设计并实现了一个基于**事件驱动微服务架构**的企业级智能招聘平台，服务之间通过 **NATS JetStream** 消息队列进行异步通信，并采用**事务发件箱模式**确保了分布式系统中的数据最终一致性。

Cloud-Native Image Annotation System

2025 年 2 月 – 2025 年 6 月

云架构师 & 全栈开发者

- 设计并实现了一个基于 **AWS Serverless** 与**事件驱动架构**的图像处理系统，通过并行的 Lambda 函数异步调用 Google Gemini API，将用户上传图片的 **API 响应时间稳定在 500ms 以内**。

个人简介 (Personal Profile)

我是一名追求全面发展的复合型工程师，拥有扎实的计算机科学背景与前瞻的产品架构思维，并坚信最卓越的产品诞生于技术深度与人文关怀的交汇点。我的思维模型受益于长期且跨学科的系统性阅读，构建了以**计算机科学为深度，人文社科为广度**的 “T 型知识结构”。这让我不仅能从第一性原理理解技术，更能洞察用户和商业的本质，在解决复杂问题时找到创新性的解决方案。我渴望能将我对技术的热情、对产品的好奇心和跨学科的思考力相结合，共同打造能为用户和社会创造巨大价值的产品。