Yuan Xu

2313072@mail.nankai.edu.cn (+86) 18918007833 Github Tianjin, China

Education

Nankai University, Tianjin 2022 – 2026

Bachelor's Degree Statistics and Data Science

Work Experience

JiOu Cloud | AIAgent Full-Stack Development Intern 2025.04 – 2025.07

Project Description:

Participated in the development and performance optimization of the company's core products SalesAgent and RAG intelligent Q&A system. The project aims to build an enterprise-level intelligent assistant with contextual memory and multi-turn dialogue capabilities, serving sales automation and customer support scenarios.

Key Contributions:

- Established a comprehensive regression testing framework for SalesAgent, developed automated scripts covering core workflows and edge cases, successfully identified and resolved 8 critical defects.
- Optimized the Memory module by experimenting with three retrieval mechanisms - time-weighted embedding, query-based clustering, and dynamic summarization, achieving an 18% improvement in long conversation consistency.
- Enhanced the RAG (Retrieval-Augmented Generation) system by redesigning document chunking strategies (semantic + structural perspectives) and upgrading from sparse retrieval to hybrid retrieval (BM25 + vector fusion), improving Q&A accuracy by 15%.
- Addressed challenges such as context drift and information redundancy in multi-turn conversations, balancing accuracy, response speed, and memory usage to determine optimal retrieval strategies.
- Improved information recall rate of long-document RAG modules by constructing high-quality chunks while maintaining semantic integrity.

Dify Open Source Platform | LLM Application & RAG Ecosystem Intern 2025.04 – 2025.07

Project Description:

Participated in the plugin and data integration ecosystem construction of the Dify open source platform, focusing on Agent toolchain, RAG data pipeline, and graph retrieval capability iteration to improve usability and scalability in enterprise scenarios.

Key Contributions:

- Developed and published community plugins for v2ex, linuxdo, and GitHub, standardizing query parameters and authentication, supporting keyword search, topic aggregation, and rate limiting for real-time Agent retrieval and summarization.
- Designed and implemented Azure Blob and OneDrive data source integration with incremental synchronization, supporting OAuth/key dual-channel authentication, directory and permission mapping, chunked upload, and resumable transfer for vectorized ingestion.
- Maintained and optimized the Pinecone vector retrieval plugin, supporting batch writing, namespaces, metadata filtering, concurrency and retry strategies, improving recall quality and latency stability through enhanced chunking and deduplication.
- Engineering entity relationship modeling and Cypher query paradigms using Neo4j, exploring Graph-RAG solutions for evidence retrieval and answer generation based on relational constraints.

Project Experience

PeerPortal - Decentralized Study Abroad Information Platform July 2025 – Present

Full-Stack Developer Code

Project Description: A decentralized bilateral information platform for study abroad that integrates intelligent dialogue, real-time messaging, forum communication, file management, and precise matching. The project adopts a modern full-stack architecture to provide comprehensive and personalized study abroad application guidance services for applicants and mentors, breaking traditional information barriers.

Technical Architecture:

- **Frontend:** Built modern SPA based on Next.js 15 + React 19 + TypeScript, implemented responsive design using Tailwind CSS + Radix UI component library, utilized Zustand for state management.
- **Backend:** Adopted FastAPI + Python to build high-performance asynchronous API services, integrated Supabase (PostgreSQL) as the main database, supporting WebSocket real-time communication.
- **Infrastructure:** Docker containerized deployment, supporting enterprise-level features such as file upload storage, user authentication, and database optimization.

Key Contributions & Responsibilities:

- Independently designed the full-stack system architecture, built a comprehensive platform containing 9 core business modules, covering user management, forums, real-time messaging, file upload, and intelligent matching.
- Developed modern SPA frontend using Next.js 15 App Router, implementing user authentication, profile management, forum discussions, and real-time chat, achieving 60% component reusability.
- Built robust backend with 50+ RESTful API endpoints using FastAPI, covering user, forum, messaging, and file upload functionalities, maintaining API response times under 100ms.
- Designed 21-table database architecture, improved database response time by 30% through indexing strategies and query optimization.
- Implemented WebSocket-based real-time communication system, supporting one-on-one mentor-student chat, message status management, and online presence, with latency under 50ms.
- Designed three-tier role-permission-resource model, implemented fine-grained API access control through JWT stateless authentication and middleware, ensuring data security and privacy.
- Supported 1000+ concurrent connections, achieved 99.9% message delivery success rate through asynchronous programming, connection pool management, and message queuing.

NKUWiki Knowledge Graph Construction (LLM-Driven Campus Q&A) October 2024 – Present

ETL Architect & Security Engineer Code

Project Description: A Nankai University campus knowledge sharing platform based on RAG (Retrieval-Augmented Generation), built with the philosophy of "Open Source · Co-governance · Universal Benefit" to create a Nankai knowledge community. The project processes 10GB+ data, covering 3000+ posts from Xiaohongshu, Zhihu, and campus marketplace, achieving 96% data cleaning accuracy and 60% improvement in structured processing efficiency. Optimized RAG Q&A based on LangChain, built 500k+ token knowledge base, achieved 91% retrieval recall rate with HuggingFace embedding models, optimized Q&A response time to 1.2s with 82% accuracy.

Key Contributions & Responsibilities:

- Designed and implemented a distributed crawler system supporting heterogeneous data sources, utilizing Playwright + Selenium for modern SPA and traditional web scraping, improving complex DOM processing performance by 40%.
- Conducted in-depth reverse engineering analysis of campus marketplace, cracked API signature algorithms (HMAC-MD5, timestamp verification), reverse-engineered JavaScript encryption algorithms to implement dynamic signature generation, achieving 95% data acquisition success rate.
- Implemented dynamic User-Agent rotation, IP proxy pool management, and request frequency control, achieved 99.5% crawler uptime and daily collection of 100k records through browser fingerprint obfuscation.
- Designed asynchronous ETL pipeline integrating Qdrant vector database, Elasticsearch, and MySQL, implemented incremental data synchronization with processing latency under 100ms.

Technical Skills

Frontend Development

- Proficient in modern frontend technology stack: React 19, Next.js 15, TypeScript 5, with capability to build large-scale SPA applications.
- Skilled in UI frameworks: Tailwind CSS, Radix UI, MUI, expert in responsive design and component-based development.
- Master state management: Zustand, Redux; frontend engineering tools: Webpack, Vite, Turbopack

- Familiar with Docker containerization technology and related ecosystem, capable of independently completing service packaging, deployment, and maintenance.

Backend Development

- Proficient in Python, utilizing FastAPI to build high-performance, asynchronous RESTful APIs and WebSocket services.
- Master database technologies: PostgreSQL, MySQL, with capabilities in database design, query optimization, and indexing strategy formulation.
- Familiar with cloud service platforms: Supabase, Vercel, as well as Docker containerization technology and CI/CD deployment processes.
- Understanding of other backend frameworks like Django, Spring Boot, with cross-technology stack development capabilities.

AI Application Development

- Master LangChain framework, with independent capabilities in designing and developing RAG (Retrieval-Augmented Generation) and Agent applications.
- Familiar with agent reasoning frameworks such as ReAct and Self-ask, understanding core mechanisms of Agent state management, tool invocation, and multimodal interaction.
- Possess practical experience in Prompt Engineering, skilled in designing, optimizing, and evaluating prompts to elicit optimal model performance.
- Understanding of model fine-tuning processes, with experience using platforms like ModelScope (SWIFT framework for LoRA fine-tuning) for model training.
- Have experience in frontend-backend integration of AI applications, including technical implementations of streaming responses and real-time interaction.