刘浩 ♂

On job, seeking for new job · 38 · Master · 14 years experience

13564496085 593954214@qq.com myhomestar

Self-description

With over 13 years of experience in the financial IT industry, I possess strong Java backend development skills combined with hands-on expertise in large language models (LLM) applications. I am proficient in microservices architecture design and optimization, system upgrades, and containerized deployments, with extensive experience using Spring Boot, Spring Cloud, Hibernate, MyBatis, Docker, Kubernetes, and OpenShift. I have a long-term focus on natural language processing (NLP), generative AI, and large language model projects, skilled in Python, PyTorch, TensorFlow, Retrieval-Augmented Generation (RAG), LoRA fine-tuning, and reinforcement learning techniques. I have successfully delivered multiple enterprise-level AI tools, such as Smart Jira and the Automated Content Review platform (ACR), significantly improving business efficiency and driving intelligent transformation.

Fluent in English with strong communication skills, able to understand diverse accents. Experienced in leading teams from zero to delivering innovative AI projects, demonstrating strong responsibility, resilience under pressure, and a continuous passion for learning and technology advocacy.

My personal CSDN blog has over 1 million visits (1,062,343+), with 302 original technical articles published: https://harryliu.blog.csdn.net/

Work Experience

Citibank Financial Information Services (China) Co., Ltd **2014/06-To Present**

Java Senior Software developer

1. Built on a solid foundation in Java Web backend development, and progressively transitioned into AI technologies, with a focus on Natural Language Processing (NLP), deep learning, and Large Language Models (LLMs). Successfully led the design and implementation of several enterprise-grade AI projects.
2. Served as the leader and instructor of the AI Interest Group in the MOT department (2023–2025), mentoring team members from the ground up in LLM and NLP applications. Successfully incubated several award-winning projects, including "Smart Jira", "BizProb Bot", and "Smart Email Finder".
3. Developed a text auto-formatting system leveraging LLM platforms such as Gemini and Azure OpenAI, featuring grammar correction, case transformation, and abbreviation standardization. The system significantly reduced manual effort and improved data consistency.
4. Built a document content auto-update engine using Gemini 1.5 Pro and Azure OpenAI. It intelligently rewrites structured content based on newly input text, greatly reducing the need for manual editing.
5. Proficient in major LLM platforms (Azure OpenAI, Google Gemini, LLaMA3) and fine-tuning techniques. Delivered enterprise-grade LLM APIs using Python and FastAPI, enabling smart transformation of legacy business processes.

Key Achievements:
1. Led the MOT department in building AI capabilities from scratch and successfully launched multiple production-grade AI tools.
2. In 2024, led the ” Smart Jira “ project, winning 1st Prize in the company's GenAI Innovation Competition (¥5,000).
3. In 2022, developed “Smart Email Finder”, winning a total of ¥12,000 in the internal Innovation Champion competition (Rounds A & B).
4. In 2021, developed the Raspberry Pi-based AI car project "123 Freeze Game", winning 1st Prize in the Unit2 department tech contest.
5. In 2010, designed an automatic email form reply system, winning ¥12,000 across Rounds A & B of the company’s Innovation Champion competition.

Citibank Financial Information Services (China) Co., Ltd **2014/06-To Present**

NLP

1. Led core financial system modernization by migrating legacy WebLogic-based services to Spring Boot microservices. Adopted Netflix OSS stack (Eureka, Feign, Zuul) to enable resilient service governance.
2. Spearheaded containerization initiatives, successfully Dockerizing dozens of services and deploying them on Red Hat OpenShift with hybrid zero-downtime rollout.
3. Oversaw platform-wide upgrades: Spring Boot 2→3, Hibernate, Spring Cloud, JDK8→17. Modernized communication stack by replacing Tibco Queue with Kafka and Hessian with REST APIs.
4. Developed Watch Tower, an internal OpenShift monitoring platform enabling real-time pod management and auto-scaling across environments.
5. Designed and delivered an Oracle data archival solution leveraging FastAPI and Shell for automated cleanup, significantly improving query performance.
6. Actively drove AI adoption, with hands-on experience in LLMs, RAG, and prompt engineering. Delivered tools such as Smart JIRA and AI-powered internal bots.

Key Achievements:
1. In 2024, completed large-scale Spring Boot 2.x → 3.2.4 migration for 100+ codebases within 4 months, including full Spring Cloud stack and JDK17 upgrades
2. In 2023, migrated 50+ microservices to OpenShift, building a real-time auto-scaling mechanism based on queue metrics
3. In 2017, independently delivered the "Swap" project with zero bugs in production, earning recognition from US and China senior leadership

Pactera **2012/07-2014/06**

Java Developer

1. Assigned to Citi China to support and enhance enterprise-level financial systems.
2. Developed backend modules using "Java", "Spring", and "Hibernate", and collaborated with frontend development using "Google Web Toolkit (GWT)" for interactive features.
3. Participated in "requirement analysis, system design, unit testing", and "deployment", contributing to improved system stability and user experience.
4. Recognized for outstanding performance and successfully converted to a full-time Citi employee in 2014.

Infosys **2011/06-2012/06**

Java Developer

Participated in the full-cycle development of a supplier rating system for Volkswagen. Responsible for Java backend API development and frontend interaction design using jQuery. Gained hands-on experience with software development processes and international project collaboration standards demonstrating strong coding proficiency and technical documentation skills.

Project Experience

[Java] Watch Tower - To Monitor Cloud Services **2025/01-To Present**

Project Leader

Description

To improve operational efficiency across Dev, UAT, and PROD environments on Red Hat OpenShift, led the development of Watch Tower, a unified visualization platform for managing microservices across multiple clusters.

The system provides a centralized dashboard displaying each service's OpenShift cluster, number of active Pods, and deployment status. It supports one-click operations such as Scale Up/Down and service location, greatly simplifying day-to-day maintenance tasks.

I was responsible for the overall system architecture, data collection logic, UI development, and backend orchestration, ensuring a stable, user-friendly platform. This solution replaced fragmented CLI-based operations with centralized visual management, significantly enhancing operational speed and accuracy.

[Java] Spring Cloud Config Platform **2025/01-To Present**

Project Leader

Description

Spearheaded the introduction of Spring Cloud Config, replacing the legacy config-content.jar-based setup with a centralized and dynamic configuration management system. Configuration

data was stored in Oracle, with Kafka used to broadcast change notifications, enabling real-time config refresh without restarting microservices.

Responsible for the overall architecture design, Config Server setup, microservices-side integration, and the definition of unified configuration templates and deployment processes. Successfully migrated 100+ microservices with zero disruption.

This initiative significantly improved configuration change efficiency and system maintainability, reduced manual error risk, and increased operational efficiency by over 50%.

[AI-LLM] BizPro Bot (Business Q&A Assistant) **2025/01-To Present**

Project Leader

Description

BizPro Bot is an enterprise-grade AI assistant powered by large language models (LLMs), designed to improve employee efficiency and accuracy in accessing business information.

Key Features:
1. Document Q&A and Summarization: Users upload internal documents, and the system uses Gemini-1.5-Pro to generate summaries and answer related questions.
2. JIRA Knowledge Search: The system fetches relevant JIRA ticket content in real-time based on user queries, builds dynamic Q&A context, and generates precise responses to cross-platform inquiries.
3. Technologies: RAG framework, Hybrid Search (MD25 + Vector Search), Cross-Encoder reranking, Gemini-1.5-Pro

Impact: Addressed fragmented information access and inefficient internal search, significantly improving operational and support response efficiency.

[AI-LLM] Automatic Narrative Text Update **2024/06-To Present**

Project Leader

Description

This system automates processing and updating narrative texts from upstream sources, replacing the inefficient manual conversion of all-caps text into customer-friendly formats. Using Python's difflib to accurately extract text differences, the diff along with the previously formatted client text is fed into GenAI models (e.g., Gemini or Azure OpenAI) to automatically generate updated client text reflecting the latest changes. The fully automated workflow reduces manual effort, improves text consistency, and enhances delivery efficiency. The solution is successfully deployed in daily event handling, significantly advancing customer communication automation.

[AI-LLM] ACR – Automatic Code Reviewer **2024/06-To Present**

Project Leader

Description

Designed and developed an internal Automated Code Review platform (ACR), integrated into the Bitbucket Pull Request (PR) workflow. Leveraging large language models (such as Gemini/GPT-4), the platform automatically analyzes code changes and generates structured review suggestions, reducing manual code review workload and improving code quality.

Key features:
1. Automated code change retrieval: Automatically fetches code diffs from Pull Requests via Bitbucket API.
2. Intelligent review suggestion generation: Utilizes large language models to perform semantic analysis on code diffs and generate targeted improvement suggestions.
3. Automated comment submission: Uses Bitbucket Comments API to post review suggestions as comments directly on the PR page.
4. Real-time trigger mechanism: Supports Webhook-triggered review processes for instant feedback upon PR creation or update.

Enterprise Business Q\&A Bot with LoRA Fine-tuning and PPO Reinforcement Learning **2024/03-To Present**

Project Leader

Description

Developed an intelligent business question-answering bot based on internal JIRA ticket data using the LLaMA 3-8B model. The project employed a multi-stage training pipeline: first, leveraging high-quality question-answer pairs generated by large language models to fine-tune LLaMA via LoRA adapters, enabling the model to acquire core business knowledge. Subsequently, reinforcement learning with PPO was applied to optimize response accuracy and interaction quality based on real-time user feedback. An incremental training mechanism automatically detects and incorporates newly created JIRA data to keep the model's knowledge up-to-date. The system also integrates Retrieval-Augmented Generation (RAG) techniques, combining vector search with generative responses to fill knowledge gaps and enhance answers to complex business queries. Adapter version management ensured stable training workflows and continuous knowledge accumulation. The entire training and deployment process was highly automated, delivering a scalable and reliable AI-driven Q\&A solution.

[Java] Tibco Queue Migration & REST API Refactor **2023/07-To Present**

Project Leader

Description

Led a team of 4 to upgrade two key communication architectures, ensuring business continuity and timely releases. Successfully migrated asynchronous inter-service messaging from Tibco Queue to Kafka, enhancing throughput and message observability. Simultaneously transitioned synchronous calls from Hessian to standard Spring REST API, resolving Hessian's compatibility issues with JDK 17 and Jakarta EE.

Lightspeed – Automated CI/CD Framework Framework Designer **2023/05-To Present**

Project Leader

Description

Lightspeed is a webhook-driven automated CI/CD system purpose-built for Bitbucket-based development workflows.

On each code push to a Bitbucket repository with an active webhook, the system automatically triggers a Jenkins pipeline, executing the full delivery lifecycle: Maven build, SonarQube code quality checks, Docker image creation, and UDeploy-based deployment.

Automation tasks are implemented using standardized Groovy scripts, centrally maintained in a dedicated repository. Each pipeline execution dynamically pulls the latest scripts to ensure reusability, maintainability, and process consistency.

This project significantly improved build and deployment efficiency, reduced manual effort, and optimized the DevOps lifecycle. It is particularly effective for large-scale microservices environments requiring continuous integration and rapid delivery.

[Java] Upgrade from Spring Boot 2.x to Spring Boot 3.2.4 and jdk17 upgrade **2024/06-2024/12**

Project Leader

Description

Led a 4-person team to upgrade over 100 codebases within the ASPEN system—including 50+ microservices and numerous shared JARs—without disrupting business operations. Key achievements:
1. Upgraded Spring Boot from 2.6.2 to 3.2.4, Hibernate to 6.x, and Spring Cloud to 2023.0.5.
2. Migrated core components: replaced Netflix Gateway with Spring Cloud Gateway, and Netflix Feign with Spring Cloud OpenFeign.
3. Designed upgrade plans, performed compatibility analysis, adapted critical dependencies, and ensured smooth system transition with improved security and maintainability.
Also led a JDK upgrade initiative, migrating 50+ microservices from JDK 8 to JDK 17, resolving all compatibility issues and preparing the system for long-term evolution.

[AI-LLM] Smart JIRA **2024/06-2024/12**

Project Leader

Description

Led the development of Smart Jira, an intelligent tool built on Gemini-1.5 Pro / 2.0 Flash to automate Jira ticket creation. It parses user inputs (text, voice, images, emails) and generates structured issue fields (Summary, Description, Acceptance Criteria, Story Points, Issue Type). Integrated with the Jira API, users can create issues with one click, streamlining agile workflows.

Key Highlights:

1. Uses LLM 1to interpret user requirements and generate standardized Jira tickets
2. Supports multiple input formats (text, audio, images, email)
3. Seamless Jira API integration for automated issue creation
4. Improves task creation speed and accuracy, enabling agile efficiency

[AI-LLM] Text Format – Canary (Corporate Action Narrative Refinery) 2024/01-2024/06

Project Leader

Description

An AI-powered system for automatic refinement of narrative text, handling case conversion, grammar correction, formatting, and acronym standardization. Integrated into the ASPEN platform, it streamlines the processing of client-facing descriptions in corporate action workflows.

Key Achievements:

1. Processes ~35 texts per day, saving ~19 minutes per item
2. Significantly improves consistency and reduces manual effort
3. Provides a scalable template for future intelligent automation in corporate action processes

[Java-Cloud] Microservices Migration to OpenShift 2023/06-2024/05

Project Leader

Description

Led a team of 5 in migrating enterprise-level microservices from traditional Linux/VM deployments to a private OpenShift cloud. The legacy system included multiple VM/Linux environments. The migration involved containerizing each microservice and implementing automated pipelines with LightSpeed (including code security checks, Maven builds, Docker image creation, and UDeploy deployment) for deployment on the OpenShift platform, ensuring each service runs independently.

To ensure business continuity, designed a hybrid deployment strategy supporting dual environments (Linux + OpenShift) during migration, with new projects fully deployed on the

private cloud.
Responsibilities included overall migration strategy, container build script design, CI/CD process refactoring, and deployment validation, achieving zero production incidents throughout the transition.

[Java -Cloud] OpenShift Pod Auto-Scaling System **2023/02-2023/05**

Project Leader

Description

Developed an automatic elastic scaling solution based on real-time monitoring metrics such as Tibco queue size, Kafka message volume, memory usage, CPU utilization, and time windows. The system dynamically adjusts the number of Pods for specific microservices by invoking OpenShift CLI commands and Kubernetes SDK, enabling automatic scaling according to business load.

This solution significantly improved resource utilization, ensured high system availability and responsiveness, reduced manual operational intervention, and realized intelligent elastic management of microservices.

[Java-Microservices] Legacy System Migration to Microservices Architecture **2018/02-2019/03**

Project Leader

Description

Led the transformation of WebLogic-based monolithic services into standalone Spring Boot microservices with embedded Tomcat, significantly improving scalability and fault isolation. Integrated Netflix stack (Eureka, Feign, Zuul) to enable service discovery, declarative inter-service calls, and unified API gateway, greatly enhancing flexibility and deployment efficiency.

Education

Fudan University 985 211 Double First-Class **2019/06-2022/06**

Software Engineering

Master

Yan'an University **2007/06-2011/06**

computer science and technology

Bachelor

Languages

English（Mastery）；Mandarin