Guangkai Ren

Algorithm Engineer

📧 Email: kevin.r5009@gmail.com 📱 Phone: +86 184 0124 5270 📍 Location: Beijing

Personal Profile

With eight years of experience in algorithm research and engineering, I specialize in the application and optimization of model algorithms. I have an in-depth understanding of various large models and classic small models, and can precisely analyze their strengths and limitations. I can efficiently optimize and implement them based on business needs, driving the innovative application of algorithms in practical scenarios.

Professional Skills

Python Java JS/TS PyTorch LLM/VLM

Multimodal Understanding and Generation Large Model Pre-training Agent

Information Retrieval/RAG Efficient Fine-tuning PEFT Contrastive Learning

Reinforcement Learning Distillation/Quantization Anomaly Detection

Model Engineering Deployment Distributed/Microservices DevOps/MLOps

Work Experience

Beijing Block Imagine Digital Technology Co., Ltd.

June 2023 - Present

Senior Algorithm Engineer

- Responsible for the development of AI agents, large model applications, and intelligent search systems, driving the implementation of AI technology in blockchain, cryptocurrency trading, and information retrieval.
- Led the architecture design and technical implementation of multiple AI projects, including LLM-based agents, smart contract interactions, and automated trading systems.
- Participated in the development of blockchain ecosystems, building AI-driven DeFi interactions, cross-chain asset management, and automated smart contract execution solutions.
- Responsible for the design and optimization of AI search engines, enhancing the parsing capabilities for complex queries and implementing multi-agent collaborative information retrieval systems.
- In the field of AI-based quantitative trading decisions, combined LLM and reinforcement learning to optimize trading strategies and improve the accuracy of market analysis.

China National Knowledge Infrastructure Technology Co., Ltd.

November 2016 - June 2023

Research and Development Engineer & Senior Project Manager

- Responsible for the design and development of search engines and recommendation systems, including full-text retrieval, multi-route recall, and optimization of personalized recommendation algorithms.
- Led the development of a text correction system, combining NLP technology with multimodal feature fusion to enhance correction capabilities in complex scenarios.
- Designed and implemented an AI-enhanced retrieval (RAG) system, optimizing vector retrieval and semantic understanding capabilities to improve information acquisition efficiency.
- Responsible for the development of industrial computer vision systems, focusing on zero-shot learning solutions for anomaly detection and defect identification.
- Participated in AI model training and optimization, covering the application of large models such as BERT, Transformer, and CLIP in multiple fields.
- Constructed big data processing platforms and AI infrastructure to support offline batch processing and real-time stream processing of massive data.

Beijing Gome Online E-Commerce Co., Ltd. December 2014 - August 2016

Research and Development Engineer

- Responsible for building a product search engine based on Elasticsearch, optimizing index structures and tokenization strategies to significantly improve search response speed and result accuracy.
- Led the development of a personalized recommendation system, using collaborative filtering and matrix decomposition algorithms to precisely match users with products, effectively increasing conversion rates.
- Participated in the construction of a distributed computing architecture, utilizing Hadoop and Spark for efficient data processing, and combining Redis to accelerate popular product recommendations, significantly improving system performance and efficiency.
- Responsible for introducing Kafka to build a real-time data stream processing framework, dynamically adjusting recommendation results based on user real-time behavior to enhance user experience.
- Led the development of machine learning models, using scikit-learn and MLlib to implement logistic regression and random forest recommendation models, significantly improving the generalization and stability of the recommendation system.

Beijing Dawn Infinite Technology Co., Ltd. September 2014 - December 2014

Research and Development Engineer

- Responsible for implementing skin beautification functions based on skin color detection and bilateral filtering algorithms, significantly improving the aesthetics of portrait photos by adjusting the hue of skin areas and smoothing skin texture.
- Led the development of face-slimming and big-eye effects, using liquidation algorithms combined with dlib to extract facial feature points and precisely adjust facial proportions to achieve more desirable selfie results for users.
- Participated in the implementation of filter functions, using convolutional neural networks (CNN) and neural style transfer algorithms to provide users with diverse and artistic photo style conversion effects.
- Led the design and implementation of photo restoration functions, using deep learning algorithms to enhance the quality of low-quality photos, effectively improving issues such as blurriness and noise, and enhancing the overall visual effect of the photos.

Project Experience

Avalanche Agent MCP Framework

January 2025 - Present

Project Overview: Responsible for the development and maintenance of the Avalanche Agent MCP framework on the Avalanche blockchain network. This framework enables AI agents to automate various on-chain operations, including token trading, NFT management, DeFi interactions, cross-chain bridging, etc. The project aims to simplify the interaction between developers and the Avalanche ecosystem and promote the application of AI agents in the blockchain field.

- Developed the Avalanche Agent MCP framework, providing developers with a convenient tool for interacting with Avalanche smart contracts.
- Implemented the automation of dozens of on-chain operations on Avalanche, significantly improving development efficiency.
- Integrated LangChain and Vercel AI SDK to build AI agents that support natural language interaction and autonomous decision-making.
- Integrated with multiple mainstream Avalanche DeFi protocols to enrich the functionality of the toolkit.
- Integrated with Wormhole and deBridge to establish asset transfer channels between Avalanche and other chains.
- Developed a multi-agent system based on LangGraph, enhancing the toolkit's ability to handle complex tasks.

LLM-based Cryptocurrency Trading Agent CryptoTradeAgent

June 2024 - December 2024

Project Overview: Responsible for the design and implementation of a cryptocurrency trading agent, CryptoTradeAgent, based on large language models (LLMs). This agent can comprehensively analyze on-chain data (such as transaction records and market data) and off-chain information (such as financial news) to formulate trading strategies. By introducing a reflection mechanism, the agent can review past trading performance and optimize future decisions accordingly, achieving higher returns in the cryptocurrency market.

- Designed and implemented the overall framework of CryptoTradeAgent, including data collection, market analysis, news analysis, trading decision, and reflection modules.
- Developed the data collection module, responsible for obtaining on-chain data from platforms such as CoinMarketCap and Dune Database, as well as off-chain information from cryptocurrency financial news sources.
- Utilized LLMs to implement market analysis and news analysis agents, which are responsible for analyzing market trends and interpreting the impact of news events on the market, respectively.
- Designed the trading agent's decision logic to generate buy, sell, or hold trading recommendations based on the results of market and news analysis, and control trading volumes.
- Introduced a reflection agent based on reinforcement learning, enabling the agent to review past trading records, analyze successful and unsuccessful experiences, and provide feedback for future trading decisions.
- Evaluated CryptoTradeAgent using deepseek-r1 and GPT-o1, and compared its performance with traditional trading strategies and time-series models.

AI Search Engine Combining Parallel Query of Multiple Sub-problems and Multi-level Retrieval Strategy

July 2023 - June 2024

Project Overview: Led the development of Thought-like AISearch, an AI-based multi-agent Web information retrieval and integration system. The system mimics the human thinking process, with two core modules - planning and searching - working together to achieve in-depth parsing and efficient information retrieval for complex user queries.

- Developed the search planning agent module, which is responsible for decomposing complex user queries into multiple atomic sub-problems and constructing a dynamic reasoning graph to guide the information retrieval process.
- Developed the search execution agent module, which is responsible for executing the sub-problems assigned by the planning agent and extracting valuable information from a vast number of web pages using a hierarchical retrieval strategy.
- Introduced a multi-agent collaboration mechanism to enable information exchange and task allocation between planning and search agents, enhancing the overall efficiency and accuracy of the system.
- Designed and implemented LLM context management strategies to effectively handle long text inputs and multi-turn dialogues, avoiding information loss and interference.

Zero-shot Industrial Product Anomaly Detection System Based on Multimodal Large Model OpenCLIP

October 2022 - June 2023

Project Overview: This project constructed a zero-shot anomaly detection system for industrial defect detection. Based on the multimodal large model CLIP method, we optimized CLIP's generalization ability in anomaly detection tasks to achieve precise anomaly detection without target domain training data.

- Designed and implemented the overall architecture of the OpenCLIP anomaly detection system, including object-agnostic text prompt learning and global + local anomaly feature optimization mechanisms.
- Developed an object-agnostic text prompt learning mechanism to optimize CLIP's recognition ability from a generalization perspective, focusing on abnormal regions rather than foreground object categories.
- Implemented a global + local anomaly learning (Glocal + Local) optimization mechanism, combining global features and fine-grained anomaly features to improve detection accuracy in multiple scenarios.
- Optimized the text and visual spaces of CLIP using learnable text embeddings (Prompt Tuning) to enhance local anomaly recognition capabilities.

Literature-enhanced Retrieval System Based on Multi-route Recall and Learning-to-Rank

February 2023 - June 2023

Project Overview: Responsible for the development of an AI-based enhanced retrieval system aimed at improving the precision and efficiency of information retrieval for academic literature and corporate knowledge bases, meeting users' complex and diverse query needs.

- Participated in the design and development of the AI-enhanced retrieval system, including system architecture design, functional module development, and performance optimization.
- Designed and implemented a multi-route recall retrieval module that combines the advantages of traditional retrieval and vector retrieval to enhance retrieval accuracy and comprehensiveness.
- Developed a feature that allows users to quickly access specific paragraphs of the original literature, enabling users to bypass the literature level and directly locate to the original text fragments, improving the efficiency of literature research.

- Designed and implemented a flexible retrieval input function that supports natural language input and voice input, with capabilities for intelligent intent recognition, error correction, and standardized retrieval guidance.
- Participated in the data processing work of the system, including literature data cleaning, vectorization processing, and the construction of a comprehensive vector data repository.

Development and Optimization of Multimodal Fusion-based Text Correction System MMSpellCheck

February 2022 - December 2023

Project Overview: This project突破 traditional text correction methods' limitations in complex scenarios by leveraging the complementary advantages of multimodal information to achieve more accurate and natural text correction.

- Participated in the design and development of the MultiModalSpellCheck system, including multimodal feature extraction, fusion strategy design, and correction model training.
- Responsible for the development of the multimodal feature fusion module, which effectively integrates text, speech, and visual information to enhance the model's ability to recognize complex errors.
- Participated in the training and optimization of the correction model, including loss function design, hyperparameter tuning, and model performance evaluation.
- Designed and implemented a rule-based post-processing module to refine model outputs, further improving the accuracy of corrections.

End-to-End Framework for Text Correction Based on BERT and Transformer

November 2016 - December 2020

Project Overview: Designed an end-to-end framework that simultaneously utilizes the visual and phonetic features of misspelled characters, minimizing their misleading impact on context to enhance the recall and accuracy of corrections.

- Utilized a lightweight Transformer to detect the positions of misspelled characters, and employed BERT in the corrector to capture the visual and phonetic features of each character in the original sentence, with the corrector directly using the original sentence as input to retain the visual and phonetic features of misspelled characters.
- Adopted a late fusion strategy, combining the hidden states of the corrector with those of the detector, aiming to eliminate the misleading impact of misspelled characters.
- Implemented end-to-end joint training to synchronize the execution of detection and correction tasks, enabling the framework to perform both correction and detection tasks simultaneously and optimize performance through joint training.

Optimization of CNKI Search and Recommendation System

December 2014 - August 2016

Project Overview: The project utilized Elasticsearch as the full-text search engine and combined Spark with Flink for large-scale data processing. It also introduced deep learning models (such as DNN, Transformer) and knowledge graph construction techniques to enhance the system's semantic understanding and cross-domain association recommendation capabilities for academic resources.

- Participated in the construction of a full-text search service based on the Elasticsearch/ELK Stack, supporting complex queries and real-time aggregation analysis, and explored the application of vector search technologies (FAISS, Milvus) in recommendation scenarios.

- Led the development and application of deep learning recommendation models, from early Wide&Deep, DeepFM to later DIN, DIEN, sequential recommendation (BERT4Rec), and multi-task learning (MMOE), effectively improving recommendation relevance and user conversion rates.
- Constructed an offline batch processing platform based on Hadoop, Spark, and Flink, and a stream processing system using Spark Streaming/Flink/Kafka Streams to achieve real-time data collection, feature extraction, and computation for massive data.
- Led the implementation of online learning (FTRL, online update framework) and real-time recommendation services, building a multi-level recommendation architecture (recall → coarse ranking → fine ranking → reranking) using the Lambda architecture, microservices, and containerized deployment (Docker, Kubernetes).
- Established a real-time monitoring system centered on Prometheus and Grafana, continuously tracking model performance metrics (AUC, NDCG, Precision, Recall, CTR, CVR, etc.).
- Driven the construction of an experimental platform (A/B testing, TensorBoard, MLflow) to automate model iteration and online debugging management.

Search and Recommendation System for Gome E-commerce Platform

December 2014 - August 2016

Project Overview: Responsible for the architecture design and optimization of an e-commerce platform's search and recommendation system, enhancing user search experience and personalized recommendation capabilities. The project utilized Elasticsearch based on Lucene as the search engine and combined the Hadoop ecosystem for large-scale data processing, while introducing collaborative filtering and matrix decomposition recommendation algorithms to improve the accuracy of product recommendations.

- Optimized the search engine by constructing a product search engine based on Elasticsearch, optimizing index structures, tokenization strategies, and search recall to improve search query response speed.
- Constructed a recommendation system using collaborative filtering (CF) and matrix decomposition (MF) algorithms to build a personalized recommendation system, optimizing user-product matching and increasing conversion rates.
- Utilized Hadoop and Spark for data preprocessing and batch computing to enhance the efficiency of large-scale data processing, and combined Redis to accelerate recommendations for popular products.
- Introduced Kafka to build a real-time data stream processing framework, dynamically adjusting recommendation results based on user behavior (clicks, collections, purchases).
- Implemented recommendation models based on logistic regression and random forests using scikit-learn and MLlib to improve the generalization capability of the recommendation system.

Education Background

Trine University (the U.S.)

May 2024 - May 2026

Master in Information System

University of Science and Technology Beijing

September 2020 - January 2023

Master of Software Engineering

Luoyang Normal University September 2009 - July 2013

Bachelor of Computer Science and Technology

Certificates and Honors

- Intermediate Software Designer (National Qualification Examination)
- Senior Project Manager (National Qualification Examination)
- CET-6, Score: 467
- Outstanding Graduate (University Level), 2023