Baiyang Qu

Website: https://baiyang66666666.github.io/ *Email: baiyangqu6@gmail.com*

Technical Skills

Programming Languages: JavaScript/Node.js, Python, NestJS, Java, SQL, C++, MATLAB

AI & Machine Learning:

- LLM Deployment, Fine-tuning, RAG, Prompt Engineering
- Azure OpenAI, Google Cloud AI Platform
- LangChain, PyTorch, TensorFlow, Hugging Face Transformers
- Image Classification, Text Processing, Time Series Analysis
- Model Evaluation & Monitoring

Cloud Computing & Tools:

- Cloud Platforms: Azure, Google Cloud Platform, Cloudflare Workers, DigitalOcean
- Databases & Caching: MongoDB, BigQuery, GCS, Redis, MySQL
- Big Data & ETL: Apache Spark, Azure Data Factory, Google Cloud Dataflow
- Data Visualization: Looker, Tableau, Power BI

DevOps:

- Microservices: NestJS, Docker, Kubernetes, Cloud-based deployment (Azure, Google Cloud)
- Azure DevOps**,** CI/CD, Version Control tools
- Rest API, Postman, Linux/Bash, PowerShell, HPC

Professional Experience

05/2024-Now VivaCity Technologies Limited AI Engineer

- ***LLM ChatBot & Multi-Functional Agent System (RAG Project):***

  Developed a distributed LLM-driven B2B ChatBot and B2C AI Agent system using Cloudflare Workers and LangChain, integrating OpenAI and Azure models to manage real-time conversations and context-aware text generation. Implemented Retrieval-Augmented Generation (RAG) technology to improve text accuracy by combining retrieval and generation methods. Designed and deployed a serverless architecture with Cloudflare Workers and Durable Objects for low-latency, global deployment. Utilized asynchronous programming and Cloudflare Queue for efficient data processing and task management. Developed AI agents for dynamic task handling and integrated LangChain for seamless model interaction, while optimizing SQL databases for efficient storage and retrieval of user data and interaction history.

  Key Technologies: Cloudflare Workers, Cloudflare Queue, LangChain, OpenAI, Azure, SQL, Node.js, CI/CD.

- ***Azure AI-based Automated Image Classification Microservice:***

  Developed an automated image classification system for a mini-program where users upload images to be categorized. The system supports automatic classification, manual categorization, and log modifications for future model training. Fine-tuned a Large Language Model (LLM) with labeled data for enhanced classification accuracy. Built the backend with NestJS and integrated Redis Queue for batch processing, while using MongoDB to store metadata and classification logs. Implemented an error correction module to improve model training and ensure continuous improvement. Achieved 91% accuracy, significantly streamlining the classification process and improving team efficiency.

  Key Technologies: Azure AI, LLM Fine-Tuning, NestJS, MongoDB, Redis.

- ***WeChat Mini-Program User Analysis & Business Insights:***

  Designed and developed a real-time dashboard to track user behaviors and query volumes for a WeChat mini-program, enabling data-driven business insights. Utilized Google BigQuery for large-scale data analysis and

Looker for real-time visualizations. Applied SQL for data cleaning and aggregation, using time-series analysis to identify trends in active users and query volumes. Provided actionable insights that helped optimize marketing strategies and enhance user engagement. Automated data updates, greatly improving decision-making efficiency.
Key Technologies: BigQuery, Google Cloud Storage, Looker, SQL, Time-Series Analysis.

- ***Industry-Specific Translation & Terminology Management System:***

  Built a custom translation system for industry-specific terminology, offering accurate translation and glossary management services. Designed and implemented a glossary datastore using Cloudflare D1 Database to manage industry terms, ensuring consistency and accuracy in translations. Developed RESTful APIs for glossary updates and integrated Azure OpenAI for customized machine translation. Utilized SQL for efficient data storage and real-time glossary synchronization, providing flexible and fast translations.
  Key Technologies: Cloudflare Workers, Cloudflare D1 Database, Azure OpenAI, SQL, RAG, RESTful API.

03/03/2023-28/07/2023 VoiceBase, Inc., A LivePerson (LPSN) Company Student Internship

- **Re-implemented MOS prediction for synthetic speech using LDNet with MobileNet and RNN/FNN decoders.**
- **Applied LDNet to VCC2018 and BVCC datasets, and analyzed performance against the original paper.**
- **Gained experience managing data and algorithms on HPC systems.**

10/06/2021-13/11/2021 ByteDance Algorithm Intern

- **Verified VR positioning accuracy and analyzed motion tracking data to improve algorithms.**
- **Collaborated with a multidisciplinary team to refine positioning systems.**

10/2017-08/2021 Little Sunflower Volunteer Team in Qingdao University Team Leader

- **Led community and rural volunteer activities, focusing on education and accessibility.**
- **Developed VR software for children's English learning and virtual travel for elderly individuals.**

Education

09/2017-06/2021 Qingdao University(81.68/100) Bachelor of Engineering

Major in Electronic Information Engineering

Modules: C Language Programming, Embedded System and Application, Signal and System, Algorithms and Data Structure, Python Programming, Circuit Principle, Object-Oriented Programming, Pattern Recognition, LabVIEW, etc.

09/2022-09/2023 University of Sheffield (Completed with Distinction Degree) Master of Science

Major in Computer Science with Speech and Language Processing

Modules: Speech Technology, Scalable Machine Learning, Text Processing, Speech Processing, Natural Language Processing, Machine Learning and Adaptive Intelligence, Team Software Project, Computer Professional Issues

Dissertation: ***Application of ASR in the prediction of Transient Loss of Consciousness consultations***

Description: This dissertation investigates using Automatic Speech Recognition (ASR) to predict the causes of Transient Loss of Consciousness (TLOC) consultations (epilepsy, functional seizures, syncope) through analyzing patient conversations.

- **Implemented attention-based methods (Whisper and Bert)** for enhanced Automatic Speech Recognition (ASR) performance.
- Extracted discriminative features from ASR for model training. **Semantic and formulation effort features are extracted from patients' clinical recordings and transcripts**.
- **Developed an ASR system integrated with classical machine learning models (Random Forest, Logistic Regression, SVM)** for automatic TLOC prediction.

Other Experience

AWARDS The ‘Outstanding Graduates of Qingdao University(2020)/ The University Academic Excellence Scholarship(2017-2021)/ The Honorable Mention on Mathematical Contest in Modeling(MCM/ICM)(2020)/ The First Prize on National Mathematical Contest in Modeling(2019)/ The First Prize on National Undergraduate Electronic Design Contest(2019)/ The First Prize on Provincial University Internet of Things Contest of Innovation (2019)