Xue Xia Junior SWE

Sunnyvale, CA • xiaxue@umich.edu • +1 (734)-489-5185 • linkedin.com/in/xuexia2001 • github.com/haasyut

Technical Skills

Languages: Python, C, C++, R, SQL, Bash, Perl, HTML, CSS
Data & Machine Learning libraries: TensorFlow, PyTorch, Scikit-learn, Hugging Face, LangChain, RAG, LLM prompting
Databases: Relational (MySQL, PostgreSQL), NoSQL (MongoDB, Redis）
Cloud Services, DevOps & Dev. Tools: Docker, CMake, Spark, DevOps (CI/CD), Cloud (AWS: EC2, S3, Lambda), Version Control (Git), Jupyter, Django, Flask
Statistical Methods: Hypothesis testing, Regression Analysis, Bayesian Statistics, Time Series Analysis, PCA

Education

University of Michigan - Ann Arbor M.Sc. in Statistics — Ann Arbor, MI — Sep. 2023 - May 2025
University of Science and Technology of China B.Sc. in Statistics — Hefei, China — Sep. 2019 - Jun. 2023

Work Experience

TikTok — Software Engineer Intern — San Jose, CA — Jun. 2024 - Sep. 2024

• Played a pivotal role within the **Software Development Life Cycle**, by writing clean and scalable code, and collaborated within a **pluridisciplinary team** including **Database Infrastructure**, **Ads Team and Video Content Team**, to eliminate bottlenecks and align goals.

• Developed a benchmark system for a **MySQL**-based vector database, integrating **Flask** to compare the performance baselines of **PostgreSQL** and **Redis**, which reduced manual validation steps by **90%**.
• Developed a scalable, object-oriented CLI tool with **Python** and **C++** for distributed ANN vector search using **NumPy**, **SciPy**, integrating modular indexing with fault-tolerant design and enabling performance benchmarking (QPS, latency, recall) in a production-like environment.

• Managed **Docker Containers**, implemented automated benchmark tests in the CI pipeline using **YAML**, and employed **Perl** for logging results into a database, integrating **GitLab CI** with container orchestration and script automation.

• Refined ANN search parameters with HNSW, IVF, L2**,** and Cosine using **FAISS**, reducing **Top-K** query latency by **10%** while maintaining recall accuracy.

• Crafted **Jenkins** Pipelines for automating the build and deployment of **Docker** Images on **AWS EC2**, improving system reliability and deployment success rate by **30%**.

• Delivered **technical presentations** on new designs and development plans, maintaining high engagement by asking calibrated questions.

Zhongxing Telecom Equipment (ZTE) — Software Engineer Intern — Nanjing, China — Jul. 2022 - Nov. 2022

• Utilized **Apache Spark** and **Scikit-learn** to extract user behavior signals from features such as click-through rate and dwell time, successfully generating relevance labels and constructing a training dataset of **1** million samples.

• Created and fine-tuned a Ranking **SVM** model in **Python** with grid search and cross-validation, optimizing hyperparameters to boost accuracy by **6%**.
• Developed a ranking model using **PyTorch** to optimize the performance of an internal search engine by focusing on data construction, model training with supervised learning, and conducting offline evaluations, achieving an **8%** improvement in precision.

• Executed offline A/B testing with historical search logs to deploy new models that incorporated optimization techniques and regression analysis, realizing a **3%** improvement in **Normalized Discounted Cumulative Gain** (**NDCG**) over various query types.

• Provided **support** and conducted **code reviews** for other developers, while remaining available and accessible to encourage collaboration.

CambioML — Software Engineer Intern (Startup Volunteer) — Remote, CA — Jan. 2023 - Mar. 2023

• Assessed **Uniflow** pipeline efficiency with **Apache Spark** and **Kafka**, using **AWS Lambda** and batch processing to speed up processing.

• Optimized **Open-Source LLM**s for structured data extraction using Domain-Specific data, **LoRA**, and **Prompt Tuning** techniques, integrated with **Uniflow** and a **Hugging Face** model to transform content into table formats and generate Q&A pairs, improving information processing efficiency.

• Crafted example workflows utilizing **LangChain**, integrating modular design and chainable operations, to demonstrate **Uniflow**'s capabilities in sophisticated document transformation scenarios like PDF-to-database and key-value extraction.

Machine Learning and Data Projects

Scalable NLP Web App: Sarcasm Detection using BERT and Streamlit — Jan. 2024 - Apr. 2024
• Built and fine-tuned a **BERT**-based classifier to detect sarcasm, achieving 92% accuracy on 30,000 labeled data from Kaggle.
• Deployed the model using **Streamlit** and **Hugging Face Spaces**, enabling real-time web-based inference with an interactive UI.

Time Series & Mechanistic Modeling: ARIMA vs SEIR on COVID-19 Forecasting — Jan. 2024 - Apr. 2024
• Conducted time series and mechanistic modeling comparison (ARIMA vs SEIR) in **R** on COVID-19 data for trend forecasting.