# YIXIAO ZHANG

**Date of birth:** 27/10/2003 | **Place of birth:** Chengdu, China | **Phone number:** (+86) 13558882230 (Home) | **Email address:** q18202429280@outlook.com | **Website:** https://kiko022.github.io/ | **LinkedIn:** linkedin.com/in/yixiao-zhang-2042ab2a3

## EDUCATION AND TRAINING

01/09/2022 – CURRENT Shenyang,Liaoning, China

**INFORMATION MANAGEMENT & INFORMATION SYSTEMS** Northeastern University

- **Relevant Courses:**An introduction to Database System(92), Probability and Mathematical Statistics (92) , Data Structures, Data analysis and data mining, Statistics, Linear Algebra, Advanced Mathematics, Business Big Data Analysis and Application.
- **Awards & Honors:**1.Business Elite Challenge Accounting and Business Case Competition — National First Prize,Jun.2024 2.Northeastern University 11th Overseas Economic Management Scholars Seminar — Outstanding Camper, Jul.2024

**Field of study** Information and Communication Technologies | **Final grade** 3.4976/5(85/100) | **Number of credits** 261ECTS

## WORK EXPERIENCE

### RESEARCH ASSISTANT – NORTHEASTERN UNIVERSITY – 19/07/2024 – Current – SHENYANG, CHINA

- Overview: Balanced MSSC (capacity/balance); Java + reproducible eval; beats K-means++.
- Designed and developed **balanced MSSC** algorithms under capacity/balance constraints to address poor initialization and local optima issues, delivering **~1.8×** faster per iteration neighbor scoring and **9–13%** lower objectives **(up to ~17%)** than **K means++ (50 restarts)** across six synthetic/real datasets **(n≤100k, k=10–50)**.
- Engineered a Java codebase with pluggable initializers, amortized **O(1) exchange cost updates** via statistics, and **tabu search operators (one move/swap)**, augmented by a **population based diversification** mechanism to avoid premature convergence.
- Established a **reproducible evaluation protocol** (fixed balance tolerance and restart budgets, **5 seeds, wall clock timing**) and ran **ablations** to isolate gains from the **fast cost updater, swap moves, and diversification.**

### R&D INTERN (PROMPT ENGINEER) – CHENGDU XIAODUO TECHNOLOGY CO., LTD. – 29/07/2024 – 28/09/2024 – CHENGDU, CHINA

- Overview: Ran a customer-service RAG from prototype to demo, improving routing, retrieval, and prompts.
- **Built a RAG routing agent** using GPT 4o and prompt orchestration to rewrite/classify queries into four workflows (KB Q&A,product Q&A, recommendation, comparison) with confidence based fallback, resulting in **+24%** routing precision.
- Implemented **hybrid retrieval in Elasticsearch** using BM25 + dense vector HNSW with bge-m3 embeddings, with a cross encoder re-ranker and rule based filters, achieving **+18%** top 5 recall and **−27%** hallucination rate in offline QA tests.
- Delivered a Streamlit-based **RAG demo** with streaming responses; stakeholder review loops were **42%** shorter.

### DEVELOPMENT ENGINEER – SHENZHEN BAIREN BIOTECHNOLOGY CO., LTD. – 06/10/2024 – 11/12/2024 – SHENZHEN, CHINA

- Overview: Bioinformatics platform; pipeline automation & containerization for Linux workloads.
- **Automated Python analysis of pipeline outputs** using Dockerized services on Linux, resulting in **+15%** success rate.
- **Designed GPT-4o prompts for JSON extraction** with schema validation and retry, achieving **90%** valid parses.

### AI DATA INTERN – MIGU MUSIC CO., LTD. – 24/12/2024 – 24/02/2025 – CHENGDU, CHINA

- Overview: Spring Festival content agent (LLM + Stable Diffusion); prompts/templates and QC datasets.
- Developed and tuned **prompts for LLMs and Stable Diffusion** (text to image/video) using batch templates and negative prompts for Spring Festival assets, resulting in **~35–40%** reduction in post editing workload
- Performed **structured labeling** for audio/video/images and **built datasets with QC checks** using Python scripts to validate metadata and file integrity, achieving **~25%** increase in annotation throughput.

### AI DATA INTERN(REMOTE) – HENAN SHUQING DATA TECHNOLOGY CO., LTD. – 27/06/2025 – 18/11/2025 – ZHENGZHOU, CHINA

- Overview: Tianchi (Bank Customer Product Subscription Prediction); end-to-end modeling pipeline.
- **Built a data preparation and feature engineering pipeline** using Pandas with data integrity validation, LabelEncoder for categorical features, and Matplotlib correlation heatmaps, resulting in cleaner training data.
- **Trained and compared logistic regression, random forest, XGBoost, and LightGBM** with scikit learn using 5 fold cross validation and targeted hyperparameter tuning, achieving **AUC 0.87, F1 0.51, and accuracy 90%** (LightGBM best).

**AI ALGORITHM INTERN – MIDEA GROUP (FORTUNE GLOBAL 500)** – 14/07/2025 – Current – SHANGHAI, CHINA

- Overview: Teachable smart-home AI across Chat & Memory Agents, focusing on memory architecture and evaluation.
- **Built user profiles and device-preference memory** using custom prompt-based dialog summarization on reset in a Chat Agent predemo with Streamlit UI, Redis for chat history, and GPT-4o streaming, to enable cross-session personalization.
- **Evaluated memory store options and selected Mem0 on-prem** using a Redis vector store with GPT-4o and bge-m3, optimizing labeled memory extraction and update prompts to stabilize persistence and reduce integration issues.
- Curated an evaluation dataset from online logs using time-ordering and showroom filtering, while mining logs via a Linux bastion and Python to detect bad cases and build confusion matrices, resulting in processing **~27,698** sessions into **2,309** scenario samples across **167** families and actionable defect buckets routed to owners.
- **Delivered the Chat Agent MVP** with multi-turn dialog, correction/clarification/rejection, and memory using a single GPT-4o pipeline with legacy fallback and few-shot integrated into the control screen, reducing **p50 latency to ~6 s**.
- **Improved speed and cost with two-stage SFT** using LoRA on Qwen 8B and token reduction, deploying an optimized on-prem model for inference and improving the internal benchmark from **10/22 → 12.5/22**.
- Set up an **LLM-for-judge** loop for fine-tuned **Qwen variants (dense and MoE, 3B–22B+)**, using GPT-OSS and Gemini as judges with GPT-5 as arbiter, to adjust prompts based on judgments and improve decision quality.

## SKILLS

### Programming & Frameworks:

Python (Pandas, NumPy, Matplotlib, scikit-learn) | R (statistical analysis, data modeling) | Java | SQL

### Generative AI & Conversational Agents

prompt engineering (few-shot, system/user prompts) | RAG pipelines | memory systems (Mem0) | distillation/SFT (LoRA) | embeddings (bge-m3) | Stable Diffusion (text-to-image/video) | cross-encoder re-ranking

### Vector Search & Infrastructure

Docker | Elasticsearch hybrid retrieval (BM25 + HNSW) | Redis (vector store) | Linux | Git | MySQL

### Web & Application Development

Flask (backend services) | Streamlit apps | Angular | Tableau

## CERTIFICATIONS

Zhejiang University, 29/08/2025
**SDG Global Summer School:Intelligent Technology Promotes Sustainable Development(3 credits)**

Shanghai Jiao Tong University, 30/09/2025
**SJTU SDG July Camp:Ocean Sustainability in a Changing Climate (2 credits)**

University of Oxford, 03/09/2025
**Digital Humanities @ Oxford Online Summer School 2025(AI in Research Libraries)**

All Universities of the German federal state Mecklenburg-Vorpommern (MV), 12/09/2025
**2025 SustainMV - The Sustainability Summer School**

Hertie School, 08/08/2025
**Data Science Summer School**

University of Maryland, 30/06/2025
**AI and Career Empowerment**

Hanyang University
**Hanyang University International School**

Revolutions in Science and Technology(95), Modern Society & Marketing(90.75), Global Entrepreneurship_Launch your startup(99.38), Korean for Beginners(100)

## HONOURS AND AWARDS

05/06/2025
**Estonian National Summer School Full Scholarship (17/500) €1,050 scholarship – Estonian Education and Youth Board**

22/04/2025
**Bucharest Summer University Partial Scholarship Selected for partial funding at 19th BSU (750+ participants) – Bucharest University of Economic Studies (ASE)**