Zexi (Jesse) Zhang

+86 13264500190 | j3ssezhang102@gmail.com | github.com/nagi-ovo | nagi.fun | Zexi Zhang

Education

Beijing University of Technology 2021/09 - 2025/06 (Expected)

Artificial Intelligence(senior at present), GPA 3.56/4.0 Beijing, China

- Main Courses: Python Programming(98), Linear Algebra(91), University Physics(99), Basic Experiments in Machine Learning (98), Natural Language Processing(90)
- Teaching Assistant for Python Programming (Fall 2024) & Natural Language Processing (Spring 2025)
- **TOELF: 94 | IELTS: 7.0 | CET6: 588**

Selected Awards

- **Kaggle Silver Medal (Top 5% teams)** 2024/02 - 2024/04

  Team Leader Kaggle Platform

  Achieved a **top 5% ranking** in a Google-hosted competition by developing average prompt templates, constructing task datasets, and employing **QLoRA fine-tuning** to enhance the accuracy of target prompt reconstruction using the Gemma model.
- **Provincial First Prize**, 2023 China Undergraduate Mathematical Contest in Modeling (CUMCM).
- **Honorable Mention**, 2024 Mathematical Contest In Modeling (MCM), COMAP.
- **Third Prize**, 2023 National College Students' E-commerce "Innovation, Creativity, and Entrepreneurship" Challenge.

Research Experience

Beijing Key Laboratory of Multimedia and Intelligent Software Technology 2023/02 - 2023/09

Research Intern Beijing University of Technology, China

- Developed MemoMusic 3.0 within the AI4Health Team, collaborated with an international research team from institutions including the University of Regensburg, Pandora, Cornell University, and UC Irvine, enhancing personalized music recommendations through contextual analysis and improving generation via music theory integration.
- Enhanced a **Transformer-based music generation framework** by training distinct models for Classic, Pop, and Yanni music. Implemented a novel method utilizing dominant melody with targeted Valence and Arousal as input, aligning output with music theory principles.
- Published and presented research at the **2023 IEEE ICMEW**, demonstrating improvements in enhancing listener emotional states and achieving higher user satisfaction.

Multimedia and Intelligent Software Technology Laboratory 2024/09 - Present

Research Assistant Beijing University of Technology, China

Research direction is class-incremental learning of **Vision Transformers** (ViT), supervised by Prof. Li Xiaoyan.

Internship Experience

Pony.ai 2024/04 - 2024/08

Software Development Intern Beijing, China

- Served as an LLM Researcher in the Service & Application team, spearheading research into utilizing OCR+LLM for an internal automated annotation system. Developed and implemented robust evaluation metrics and integrated **prompt engineering**, **model fine-tuning**, and **backend workflow design** to address complex business scenarios and high-variance data quality, resulting in a **70% increase in annotation efficiency**.
- **Fully involved in the project lifecycle**, collaborating closely with stakeholders from Supply Chain, Product Management (PM), and Project Architecture to integrate solutions and manage costs, gaining insights into the company's overall business operations.

Publication

MemoMusic 3.0: Considering Context at Music Recommendation and Combining Music Theory at Music Generation. L. Mou, Y. Sun, Y. Tian, Y. Sun, Y. Liu, **Z. Zhang**, R. He, J. Li, J. Li, Z. Li, F. Gao, Y. Shi and R. Jain. (2023). IEEE International Conference on Multimedia and Expo Workshops (**ICMEW 2023**). [Paper]

School Activity

BJUT-SWIFT Student Association 2023/07 - Present

Founder and maintainer

- Initiated and led bjut-swift, a **student organization** dedicated to knowledge sharing and technological innovation, organizing activities such as co-learning sessions for open courses from renowned foreign universities.
- Developed the **open learning resource repository** BJUT-Helper ★100+, utilizing Python scripts and GitHub Workflow to implement automated maintenance. Provided LaTeX thesis templates, PPT designs, campus network tools and other learning resources to provide students with a more comfortable learning experience.

Open-Source Contributions

Datawhalechina/LLM-Deploy Contributor

- **Key contributor** to open-source LLM deployment resources. Specialized in vLLM concurrency optimization, with focus on network communication, decoding, and distributed inference
- Provided practical model and service optimization strategies

Personal Projects ★ 550 Nagi-ovo

- 🤗**LLama-3-RLHF**: Implemented end-to-end LLM alignment pipeline from Meta's LLama-3-8B through **SFT(Supervised Fine-Tuning)**, **DPO(Direct Preference Optimization)**, to **PPO-based RLHF(Proximal Policy Optimization-based Reinforcement Learning from Human Feedback)**.
  - Developed memory-efficient training with **QLoRA** and **LoRA adapters** for Actor/Critic/Reward models mounted on SFT base model;
  - Built and trained **Reward Model** on 160K hh-rlhf dataset with **modified LLM head architecture**;
  - Implemented **distributed training** on 4×4090(48G VRAM) using **DeepSpeed** stage 1 and FlashAttention 2;
  - Fine-tuned with **PKU-SafeRLHF-30K dataset**, achieved **92% reduction in toxicity scores** (0.1011 to 0.0081) while maintaining model's instruction-following capabilities.
- **Alphazero-Gomoku**(★2): Utilizing a self-supervised learning paradigm, high-quality training samples were generated through the integration of **Self-Play** and **Monte Carlo Tree Search** (MCTS). Subsequently, a dual-headed neural architecture, integrating strategy and value networks, was designed and trained for rapid convergence by adopting a 1cycle learning rate policy and finely tuning PUCT hyperparameters. The model's performance in the 9x9 Gomoku game was significantly enhanced by optimizing the exploration-exploitation balance and temperature annealing mechanism.
- **CHSI-Converter** (★324): Automated tool that instantly converts Chinese academic credentials into English for education certification. Built with Flask and containerized using Docker for easy deployment and scalability.
- **CRAG-Ollama-Chat** (★79): Implemented CRAG (Corrective RAG) using **Langgraph**, significantly reducing LLM inaccuracies and hallucinations. Integrated external knowledge via web search tools and employed lightweight retrieval evaluator for document relevance and reliability. Compatible with OpenAI API and local LLMs via Ollama.
- **Cherno-CPP-Notes** (★127): Keynotes and insights from **modern C++** courses, including personal interpretations and opinions. Received positive feedback on forums for in-depth understanding and practical tips.

Technical Skills

- **Program Languages:** Python, C/C++, JavaScript/TypeScript
- **Frameworks and Tools:** Pytorch, OpenAI Triton, Docker, Next.js
- **AI Technics:** Natural Language Processing (Llama Series, GPT-2), Reinforce Learning(Deep Q-Learning, PPO, RLHF)

Last Updated on January 7, 2025