Last updated in September 2024

Zhiyu Gui

Hefei, Anhui, China guizhiyu@mail.ustc.edu.cn +86 18005859962 00ffcc.tech 知乎 00ffcc

Summary

Junior undergraduate student in Computer Science and Technology at the School of the Gifted Young, University of Science and Technology of China (USTC). GPA: 3.85/4.3 (Rank: 26/215). Specialized in LLM optimization, robotics, and autonomous systems development.

Education

University of Science and Technology of China (USTC) *Sept 2022 – Jun 2026*

Bachelor of Science in Computer Science and Technology(expected)

- GPA: 3.85/4.3, Rank: 26/215

Awards

Champion, RoboGame 2023 Robotics Competition *Nov 2023*

University of Science and Technology of China

- Responsible for robot vision, positioning, and communication.

Champion, Artificial Intelligence Innovation Application Competition *Sep 2023*

School of Information Science and Technology, USTC

- Solo project focusing on image semantic segmentation.

Silver Award, RWKV 2025 Ecosystem Content Collection Competition *Jan 2025*

Yuanshi Intelligence

- Developed the first RWKV backend inference framework supporting continuous-batching.

Projects

AttnInput: Context-Aware Pinyin Input Method *AttnInput*

- Enhanced Pinyin input method using large language models (LLMs).
- Responsible for backend model inference and training.

conRWKV: High-Concurrency RWKV Backend Inference Framework *conRWKV*

- First RWKV backend framework supporting continuous-batching and chunk prefill.
- Significantly reduced Time-To-First-Token (TTFT).
- Provided OpenAI-compatible API interface.

Skills

Backend Development: Python, C, C++, Cython, SQL

Frontend Development: PyQt

Libraries: PyTorch, Transformers, vLLM, FastAPI

GPU Kernel Development: Triton, CUDA, HIP

Hardware Development: Verilog, JLCEDA, Server Setup

Interests

LLM: LLM Reasoning, LLM Inference Optimization, Linear LLM Architecture Design, LLM Interpretability

Embodied Intelligence: Dexterous Hands, Visual-Tactile Sensors

Autonomous Driving: Inland Waterway Autonomous Navigation

Publications

AttnInput: Advancing Context-Aware Pinyin Input with Efficient Language Model Integration *Submitted to ACL2025*

Zhiyu Gui, *Xulei Sun*

- Enhanced pinyin input method using large language models (LLMs), achieving state-of-the-art performance in abbreviated pinyin input while significantly reducing training costs.
- Preprint available

Certificates

Kunpeng Ascend Training Camp Certificate *Jul 2024*
Huawei

Work Experience

Intern *Jan 2025 – Feb 2025*
Zhejiang Blue City Zhige Technology Co., Ltd.

- Participated in the development of autonomous navigation technology for inland new energy vessels.