ZEKAI CHEN

+1-607-262-1435 | zc542@cornell.edu | linkedin.com/in/zekai-chen | github.com/Zekai-Chen

Education

Cornell University, College of Engineering, Ithaca, NY

Master of Engineering in Systems Engineering (Robotics) GPA:4.0 Dec 2025

Relevant Coursework: NLP & LLMs, Robot Learning, Software Engineering, Autonomous Mobile Robots, Robotic MiniBot System, Integrating Human Driving in a VR Testbed, Frontiers of Computer Vision

The Chinese University of Hong Kong, Shenzhen, China

Bachelor of Electronic Information Engineering - Computer engineering (First Class Honours) May 2024

Relevant Coursework: Software Engineering, Machine Learning, Machine Intelligence and Applications, Microprocessors and Computer Systems, Computer Architecture, Operating Systems, Data structures, Database system, Digital System Design

Professional Experience

AI Application Research Intern | Futurewei Technologies Inc., San Jose, CA June 2025 – Present

- **Co-first Author,** *High-Fidelity 4x Neural Reconstruction of Real-time Path Traced Videos. ICCV-AIGENS 2025.*
- Designed a physics-consistency benchmark for fluid video generation models, aligning evaluation metrics with human perceptual realism; benchmarked against existing open-source methods and preparing a survey paper.
- Built agent systems for workflow automation: generated one-click semiconductor news reports for Marketing (cutting prep time from ~2h/day to seconds); auto-patched Cortex-A710 erratum in Trusted Firmware (TF-A) with Linaro (reducing manual patch effort by 90%); optimized OpenSSL by implementing the Camellia cipher in ARM assembly(boosted throughput by ~25%).

AI Engineer Intern | Nexa AI, Cupertino, CA Dec 2024 – May 2025

- Designed benchmarks for RAG pipelines & LLM rerankers (jina-reranker-v1/v2, jina-clip-v2, BGE, Qwen2.5-VL), enabling model selection now deployed in the Hyperlink product [https://hyperlink.nexa.ai].
- Built a cross-platform on-device LLM test-bed (Mac M-series & Windows x86) measuring latency, RAM, and power across FP16/Q8/Q4 models.
- Authored internal tutorials for llama.cpp, transformers, and lm-evaluation-harness, reducing onboarding cycle by 40 %.
- Refactored and standardized diverse open-source models (Kokoro, Wav2Vec2, Jina-v2, PP-OCR-v4), integrating them into production-ready pipelines; eliminated dynamic ops and improved single-batch inference speed by up to 18%.
- Shipped a Gradio interface (`nexa run <model> -gr`) with the Nexa SDK (4.7 k stars on GitHub), enabling one-command model demos. [https://github.com/NexaAI/nexa-sdk/tree/Zekai]
- Contributed front-end features (React/Node.js) including company landing page and CES 2025 demo UI. [https://nexa.ai]

Robot Engineer | Cornell Cup Robotics, Ithaca, NY Sep 2024 – Present

- Implemented autonomy stack on XRP & iRobot Create (EKF SLAM, RRT/PRM, potential-field local control), achieving <15 cm waypoint error in dynamic mazes.
- Re-engineered Disney BB-8 on open-source XRP and produced LEGO-compatible variant, enabling rapid algorithm benchmarking and showcased at the 2025 FIRST Robotics Showcase.

Software Engineer Intern | Walnut AI, Sunnyvale, CA July 2024 – Aug 2024

- Shipped cross-platform NFC offline transfer (32 KB) via Flutter; developed the WalnutX NFC plugin and integrated it into the Walnut app for seamless device-to-device sharing. [https://github.com/IgnitePro-AI/walnutx_nfc]

Research Assistant | Shenzhen Institute of Artificial Intelligence and Robotics for Society, China Sep 2023 – May 2024

- Built multi-agent LLM "Werewolf" simulation (MetaGPT + GPT-4o) with 6+ agent roles and 10+ simulated games, enabling analysis of emergent tactics in competitive scenarios.
- Integrated Whisper + GPT-4 on Pepper Robot, achieving real-time speech-to-response latency <2s; deployed in daily service interactions at Shenzhen Longgang Station (500+ user queries handled).

Leadership

Co-Founder & Co-president | Stanford University ASES (GBA Chapter), Hong Kong SAR Nov 2023 – Present

- Ran Asia Launchpad for 80+ students; mentored 4 startups.

Founder | CUHKSZ Alumni Association, Shenzhen, China Sep 2023 – Present

- Built network of 300+ alumni, launched Entrepreneurship & Zhejiang divisions.

Skills and Interests

- **Programming & Tools:** Python, C++, MATLAB, Git, Linux, AWS, Google Cloud, CUDA, PowerShell
- **Frameworks:** PyTorch, HuggingFace, Transformers, LoRA/PEFT, OpenCV, llama.cpp, Gradio
- **Robotics/AI:** SLAM, EKF, A*/RRT*, Potential Fields, LLM RAG, Quantization; skilled in paper reproduction, inference testing, benchmark design; self-taught Isaac Sim simulation; replicated models incl. π0, OpenVLA, Rekep
- **Compute and Deployment:** Proficient in scheduling/benchmarking on V100, A100, RTX 4090/5090; experienced in local deployment, debugging, and training of open-source multimodal models
- **Continuous Learning:** Enthusiastic about mastering emerging AI tools to expand skillset and apply them in practical projects
- **Interests:** Go game (3-dan level at age 11); 1500m National Third-Level Athlete; coffee brewing, hiking