Kevin Xiang Li

https://kevinx.li

Email: kevinx.li@outlook.com
TEL: 734-510-0189

Skills

Machine Learning: PyTorch, HuggingFace Transformers, Unsloth, Nsight, Gymnasium, Python, C++
Mobile Dev: Flutter, SQLite, Rust **AR/VR**: Unreal, Unity, C# **Web**: TypeScript, JavaScript, HTML, CSS

Education

- **Stanford University** — Stanford, CA, U.S.A
  M.S. in Computer Science; GPA: 4.0 — *2024. 9 – 2026. 6*
  - **Course Highlights**: Reinforcement Learning, Deep RL, Spoken Language Processing, Machine Learning with Graphs, Infrastructure at Scale, Computer Networking
- **University of Michigan** — Ann Arbor, MI, U.S.A
  B.S. in Computer Science, Minor in Linguistics; GPA: 3.87, Summa Cum Laude. — *2020. 9 – 2024. 5*
  - **Course Highlights**: Intro to ML, Intro to NLP, Computer Vision, XR & Society, Programming Languages, Compiler Construction, Intro to Operating Systems, Computer Security

Experiences

- **LLM Inference Workload Performance** — Santa Clara, U.S.A.
  ML Engineer at Nvidia — *2025. 6 - 2025. 9 (Expected)*
  - **Benchmarked leading VLMs on large scale GPU clusters**: Measured throughput and latency of leading open source VLMs like Llama 4 and Qwen 2.5 VL on large scale H100, H200, and B200 clusters.
  - **Analyzed VLM inference bottlenecks**: Pinpointed kernel-level performance bottlenecks with Nsight and PyTorch Profiler on SGLang and vLLM; identified perf gaps between frameworks under varying concurrencies.
  - **Contributed 2 PRs to boost VLMs in SGLang**: Doubled Qwen 2.5 VL vision prefill performance in SGLang through efficient attention backend. Enhanced profiling support and simplified vision preprocessing for Llama 4.
- **VideoMultiAgents: A Multi-Agent Framework for Video QA** — Stanford University, U.S.A.
  Researcher, in collaboration with Panasonic — *2024. 10 - 2025. 3*
  - **Designed multi-agent framework with modality-specific agents for video QA**: Enhanced video understanding by leveraging strengths of video, text, and graph modalities through multi-agent collaboration.
  - **Discovered that modality-specific multi-agent architectures benefit from structure and independence**: Showed that our Report architecture performs the best by aggregating opinions from independent modality-specific agents through an organizer agent and weighing strength of evidence from each modality.
  - **Achieved SOTA accuracy on popular video QA benchmarks**: Improved previous SOTA on Intent-QA by +6.2%, EgoSchema subset +3.4%, and NExT-QA by +0.4%.
- **LLM Hub Supporting Fine-tuning, Inference, and Evaluation** — Shanghai, China
  ML Engineer at GienTech Technology — *2024. 6 - 2024. 8*
  - **Devised Evaluations for LLMs**: Evaluated LLMs on metrics like BLEU, ROUGE, Levenshtein Distance, and LLM-as-a-Judge methods. Incorporated evaluation module into existing PoC product.
  - **Implemented Instruction Selection and Generation**: Leveraged latest techniques like CaR (Clustering and Ranking) and Self-Instruct to select and generate instructions for more efficient and performant fine-tuning.
  - **Fine-tuned LLMs on Multiple GPUs**: Utilized popular frameworks like LLaMA-Factory and Deepspeed to fine-tune 16 open LLMs on multiple GPUs.
- **Live-It: Image-to-3D Scene Generation** — Berkeley, U.S.A.
  Team Lead, UC Berkeley AI Hackathon 2025 — *2025. 6*
  - **Led a team of 3 to win at the world's largest AI Hackathon**: Developed Live-it, a project that transforms any image into an explorable 3D world by combining the strengths of video diffusion model with Gaussian Splatting. Won the Nitrode Turbo Mode Award amongst 1,400+ hackers and 350 projects.
  - **Architected a real-time image-to-3D pipeline integrating multiple models**: Orchestrated a workflow using Veo 3 for video generation and 3D Gaussian Splatting for reconstruction. Leveraged Visual Geometry Grounded Transformer (VGGT) for its fast feed-forward inference of camera trajectories and 3D point clouds, enabling an instant 3D scene preview which was then continuously refined in real-time.