Kevin Xiang Li

https://kevinx.li

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

Skills

Machine Learning: PyTorch, PyG, HuggingFace, LLaMA-Factory, Python, GGML, C++
Mobile Development: Flutter, SQLite, Rust **Web**: JavaScript, HTML, CSS, Emscripten, WebAssembly

Education

- **Stanford University** Stanford, CA, U.S.A
  M.S. in Computer Science; GPA: 4.1/4.3 *2024. 9 – 2026. 6*
  - **Course Highlights**: Machine Learning with Graphs, Reinforcement Learning, Animation & Simulation
- **University of Michigan** Ann Arbor, MI, U.S.A
  B.S. in Computer Science, Minor in Linguistics; GPA: 3.87/4.0, Summa Cum Laude. *2020. 9 – 2024. 5*
  - **Honors**: James B. Angell Scholar (5 consecutive terms of all A's), Class of 1935 Engineering Scholarship ($2000)
  - **Course Highlights**: Intro to ML, Intro to NLP, Computer Vision, XR & Society, Programming Languages, Compiler Construction, Intro to Operating Systems, Computer Security

Experiences

- **VideoMultiAgents: A Multi-Agent Framework for Video QA** Stanford University, U.S.A.
  Researcher *2024. 10 - 2025. 3*
  - **Designed multi-agent framework with modality-specific agents for video QA**: Enhanced video understanding by leveraging complementary strengths of video, text caption, and scene graph modalities through multi-agent collaboration.
  - **Evaluated different multi-agent architectures and identifed Report as strongest for VQA**: Report allows organizer agent to aggregate opinions from independent modality-specific agents and improves VQA accuracy by 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%. Paper under review for ICCV 2025.
- **On-Device NLP Library** Shanghai, China
  ML Engineer *2024. 7 - 2024. 9*
  - **Developed an efficient NLP library in C++**: Implemented efficient transformer inference with GGML for on-device use, supporting word segmentation and named entity recognition for Cantonese and Chinese.
  - **Optimized for edge devices**: Achieved 17x smaller model size and 3x faster inference compared to HuggingFace's implementation of ELECTRA Small, while maintaining comparable accuracy. Utilized a combination of model compression techniques including layer drop, knowledge distillation, and quantization for better balance between performance and size.
  - **Deployed cross-platform libraries for Web, Node.js, and Python**: Published PyPI and NPM libraries for development and production use cases. Customized CMake configs and C++ interface to build for Mac/Linux with Clang/GCC and WebAssembly through Emscripten.
- **LLM Hub Supporting Fine-tuning, Inference, and Evaluation** Shanghai, China
  ML Engineer at GienTech Technology *2024. 6 - 2024. 8*
  - **Devised Evaluations for LLMs**: Comprehensively 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 open LLMs on multiple GPUs.
- **Statically Contextualizing LLMs with Typed Holes** University of Michigan, U.S.A.
  Researcher *2023. 9 - 2024. 8*
  - **Enhanced code LLMs with static retrieval**: Leveraged semantic context and static error correction capabilities of language servers to enhance LLM code generation accuracy and stem hallucination.
  - **Boosted LLM coding performance significantly**: Static retrieval method resulted in 3.5x more unit tests passed on 5 realistic TypeScript benchmarks, compared to vector retrieval with GPT-4.
  - **Published at OOPSLA**: Research published at OOPSLA 2024 in Pasadena, California.