

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, Animation & Simulation, Computational Human Genomics
- **University of Michigan** Ann Arbor, MI, U.S.A
B.S. in Computer Science, Minor in Linguistics; GPA: 3.87/4.0 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

WORK

- **GienTech Technology** Shanghai, China
ML Engineer 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.

PROJECTS

- **Cantonese Dictionary Mobile App** Hong Kong Lexicography Limited
Mobile Developer 2021. 10 - present
 - **Designed cross-platform UI using Flutter:** Supports both iOS & Android, rated 4.7 stars on Play Store.
 - **Built fast search backend in Rust:** Real-time fuzzy search as you type, powered by Finite-State Transducers (FST) and Term Frequency-Inverse Document Frequency (TF-IDF) over 50K bilingual entries.
 - **Optimized responsiveness with local-first design:** Instantly-responsive UI powered by on-device SQLite database, synced with online PostgreSQL for use across multiple devices.
 - **Gained over 5K monthly active users:** Gained 10K+ installs on App Store & 6K+ installs on Play Store.
- **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.
- **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.