# Edward Chen

echen1246@gmail.com | 925-725-5285 | github.com/Echen1246 | eddiechen.xyz | x.com/ayocheddie

I build things end-to-end with attention to detail: mobile apps, backend systems, and ML pipelines. Currently hacking Android apps and optimizing on-device inference.

## Projects

**LLM Safety Alignment and Interpretability** | **openedai** | PyTorch, Modal, vLLM

- Applied mechanistic interpretability techniques to analyze refusal behavior in Qwen-2.5-72B's residual stream, identifying directional vectors that control model outputs without fine-tuning.
- Deployed inference endpoint on Modal's B200 GPUs with FastAPI and vLLM, optimizing CUDA kernels for real-time latency.

**Modified Transformer Architecture | smarternano** | PyTorch, CUDA, OpenRouter

- Re-engineered the nanochat architecture with custom layer depth and weight initialization, custom trained on Nvidia's Nemotron Nano dataset; project was publicly recognized by Andrej Karpathy.
- Built a synthetic data generation pipeline via OpenRouter to align the model's personality, achieving coherence comparable to 560M parameter models.

**Wasm Runtime and Compiler | silicon-JIT** | Rust, ARM64 Assembly

- Engineered a WebAssembly runtime in Rust that executes C/Rust programs on Apple Silicon, with a JIT compiler translating bytecode to native ARM64 machine code.
- Managed executable memory via manual mmap/mprotect syscalls for direct code generation.

**Mobile ML Project | murmur** | ONNX, Flutter, ML Engineering

- Built and released a mobile app that converts any PDF into human-quality audio entirely on-device, achieving 3-second load-to-playback with zero server dependencies.
- Deployed Kokoro-82M transformer TTS on Android via ONNX Runtime, forking the runtime to isolate inference on a background thread for performance.
- Built an NLP preprocessing pipeline by cross-compiling espeak-ng and writing Dart FFI bindings for IPA phonemization, implementing token-aware batch splitting against the model's tokenizer vocab, and caching per-sentence phonemes in SQLite to eliminate redundant compute at inference time.

## Education

**Arizona State University** | 3.7 GPA, Dean's List — August 2023 - Present

BS Data Science, BA Supply Chain Management, Minor in Economics
Activities: Scholars of Finance, Association of Computing Machinery, Chinese American Students Association

## Experience

**mymelo.org** | Co-Founder, Backend Engineer — April 2025 - December 2025

- Trained and deployed a custom CNN (EfficientNet-B0) achieving 95% accuracy on 30K medical image samples.
- Engineered the Android and iOS mobile app using Flutter and Dart with 3rd party auth and AWS.
- Led 50+ in-person interviews with clinics and faculty across ASU and UCSD to tailor and improve feature set and UX.
- Engineered a scalable data pipeline using AWS S3 and Lambda to support a high-volume image processing workflow for model inference and data archival.

**M.Y Intellectual Property** | Data Scientist Intern — May 2024 - August 2024

- Built Python scripts to extract and structure patent filing data from CNIPA and USPTO databases, automating what was previously manual research for attorney review.
- Created dashboards visualizing patent filing trends and competitor activity for client presentations.

## Skills

**Languages:** Python, Java, Javascript, SQL
**Frameworks & Tools:** Next.js, Django, Flutter, Tableau, Supabase, Flask, RestAPI, MCP, React
**Libraries:** Pandas, numpy, scikit-learn, matplotlib, transformers, OpenCV, PyTorch, torchvision, fastai, TensorFlow
**Open Source:** Contributor to PyTorch, ONNX Runtime