Ruisheng Zhao

San Jose, CA, USA

rathenzrs@gmail.com
(+1) 628-286-9012
ruishengzhao.com
linkedin.com/ruisheng-zhao-293105303

Profession Experience

AiNaDoctor Inc. San Jose

Senior Frontend & Mobile Engineer & AI *Apr. 2021 – Present*

(Frontend & iOS & Android, LLMs with llama.cpp, WebRTC)

- Spearheaded development of a scalable **iOS(Swift, Object-C)** and **Android(Java)** applications interfacing with **multimodal LLMs (images, voice, OCR)**, enabling asynchronous, real-time interaction for healthcare and marketing professionals. Integrated **AVFoundation, VisionKit, and CoreML** to support seamless multimodal inputs.
- Built **modular UI** and **feature libraries** for **iOS(Swift, Object-C) and Android(Java)** with shared WebView components reused in **React and Flutter** shells. Encapsulated core features—chat, media input, async flows—into cross-platform modules, enabling consistent UX and faster releases. Developed a reusable WebView bridge for seamless LLM integration across clients.
- Developed modular customer support interfaces using **React**, enabling real-time chat, ticket tracking, and LLM-assisted responses across web and mobile platforms.
- Collaborated closely with **product managers**, **designers**, and **ML engineers** to prototype, iterate, and launch **user-centric features**; drove rapid experiments with **A/B testing** and UX feedback loops to enhance engagement.
- Ported and optimized `llama.cpp` for on-device inference (iOS & Android) with **Metal, NDK, SIMD**; supported Q4_K_M/Q5_K quantized GGUF models, achieving **2× faster** inference on Apple Silicon and Android GPUs.
- **Tech Stack:** `React, Flutter, JS, React Native`, `llama.cpp`, GGML, Metal, Core ML, NDK, Swift, Java, Obhect-C, PostgreSQL, Prisma, REST APIs.

So-Young Technology Co., Ltd. ***Director of iOS Development*** **San Francisco** *Dec. 2017 – Apr. 2021*

- **Led strategic iOS architecture initiatives** across multiple product lines, delivering reusable components adopted in 6+ apps, **cutting app size by over 28% (140MB to 100MB)**, and **maintaining a crash rate under 0.02%** through implementation of real-time monitoring systems (OneAPM, Firebase), setting new performance and reliability benchmarks company-wide.

Anjuke (58 International) ***iOS Developer*** **San Francisco & Beijing** *Mar. 2015 – Nov. 2016*

- Led development of high-traffic financial service modules for iOS apps with **over 5 million active users**. Achieved **35% improvement** in startup speed through advanced optimization techniques such as code splitting and lazy loading, **earning "Top Performer" recognition** in 2018.

US Patents

- **4D TOOTH MODELING SYSTEM AND METHOD FOR MODELING TEETH IN 4D VIA MOBILE DEVICE-BASED RGB VIDEO AND OPTICAL SENSING TECHNOLOGY**

Publications

- **MovePose: A High Performance Human Pose Estimation Algorithm on Mobile and Edge Devices**
  Artificial Neural Networks and Machine Learning – ICANN 2024
  Lightweight CNN for real-time pose estimation on edge devices using **heatmaps**, **PAFs**, and **MobileNet**; optimized via **quantization** and **TFLite**.

Education

Yantai Nanshan University | *Bachelor in Electronic Information Engineering Sep. 2011 – Jul. 2015*

Skills

- **Mobile & Platform:** Swift, Java, Objective-C, Flutter, React Native, Metal, React, Java Script, Core ML, LiveKit Agents
- **AI & LLM Systems:** LLAMA.CPP, Prompt Engineering, Dialogflow CX, LLM Orchestration, Vision-Language Models (OpenCLIP), FAISS Vector Search, STT–LLM–TTS Pipelines, Real-Time Inference (Edge & Cloud)
- **Frameworks & Tools:** PyTorch, TensorFlow, Huggingface Transformers, React.js, Next.js, Flutter, React Native, Django, Node.js, REST APIs
- **Cloud & Data:** Google Cloud Run, AWS, Docker, PostgreSQL, Prisma, MongoDB, MySQL, Airtable
- **Languages:** Python, JavaScript, TypeScript, Java, C/C++, SQL, LaTeX