

Ruby (Yuru) Xu

Software Engineer

rubyxyrain@gmail.com • Munich, Germany

linkedin.com/in/yuru-x-38914a302

https://rubyxu.com

PROFESSIONAL SUMMARY

As a proficient software engineer and tech lead specializing in AI, machine learning, backend, Cloud, blockchain and 3D, I bring over 8 years of experience in developing advanced, efficient AI and full-stack solutions. I am adept at applying cutting-edge AI techniques to solve complex problems, demonstrating a strong commitment to innovation and leadership in team collaboration.

WORKING EXPERIENCE

X-LAB

Munich, Germany

Project Manager & Requirements Lead, Software Engineer

2026.1 - Present

Served as the project and requirements communication lead for Kühne's digital transformation initiatives, coordinating alignment between client stakeholders and development teams, driving business process analysis and project execution, and enabling the transition of core business scenarios from manual operations to systematic, scalable digital solutions.

GMI Cloud

Mountain View, CA

Large Language Model, Software Engineer

2025.5 - 2026.9

Specialized in building the routing, monitoring, and performance-tuning layer of our inference-serving infrastructure — from Python-based tooling to the Go control plane of a K8s-native global GPU platform.

Project Perflab

- Led the "Perflab" project, delivering a demo in two weeks using FastAPI and Node.js to monitor online model health and performance, capable of detecting endpoint failures within one minute.
- Integrated Preflab with our internal inference engine to automate deployment configuration tuning, systematically testing and determining the optimal parameters for service frameworks (such as vLLM, SGLang) and environment variables. It reduces deployment, tuning, and performance testing time from over 2 hours per model to less than 30 minutes.

Project Proxy-Server

- Led traffic governance and reliability for a multi-vendor LLM proxy/gateway: dynamic routing, health scoring, automated probing, failover, timeout/retry, and observability, built on Nacos, Redis, and OpenTelemetry.
- Re-architected the routing and scoring intelligence from the Python proxy into a Go control plane: built an endpoint scoring engine with a config-driven tier classifier and traffic weighting, and designed a latency-based capacity detector that catches endpoint overload dynamically.

CERBO AI

Mountain View, CA

Large Language Model, Software Engineer

2024.5 - 2024.9

Use python to develop a robust LLM agent framework for desktop applications, enhancing compatibility with mobile applications by optimizing end-side model performance.

LLM and Blockchain Engineer

- Optimized Mobile LLM Agent: Developed a highly customizable personal assistant by fine-tuning LLM models within a PC-based framework, ensuring readiness for seamless mobile integration.
- Open-Source Project - Federated Learning with Web3 Integration: Led secure and efficient individual model training through federated learning, bolstered by Web3 integration to record every model weight upload.

TECH MAHINDRA (VENDOR FOR GOOGLE X)

Shanghai, China

Machine Learning, Software Engineer

2022.6 - 2024.5

As a vendor engineer in Google X (Mineral.ai), I am proficient in Google infrastructure and development tools, such as Google3, XManager, Beam/Flume, Vertex AI, App Engine, Big Query, Cloud Run, Kubernetes, etc. Due to my large impact in project BerryQA, particularly the ML modeling, I was promoted to the tech lead and project leader of Project Dolphin, a foundation model native MLOps system, with smart labeling features to accelerate image annotations by 10x through foundation models (OWL-ViT).

Project Dolphin, Project and Tech Lead

Provided technical leadership for a team of 5 (3 frontend, 2 backend developers) to build a web image detection and classification application from the ground up and successfully launched Minimum Viable Product (MVP) within 6 months, notably improving image annotation efficiency by 10x.

- Developed backend systems with emphasis on machine learning models. By combining the foundation model with downstream tasks, the response speed was significantly accelerated while also enhancing the accuracy of the results.
- Built docker custom containers to serve large computer vision foundation models for few-shot/one-shot object detection inference. Leveraged Vertex AI to productionize and serve models via APIs.
- Designed and optimized cloud native architecture to enable scalable and resilient systems. Formulated hypotheses and performed controlled experiments to guide architectural and design decisions with empirical data. Quantitatively evaluated alternatives to determine optimal technical solutions.

Project BerryQA, ML and Backend Engineer

Developed backend features using Flask, BigQuery, App Engine to enable enhanced quality inspection model integration.

- Led efforts to advance image detection and classification models. Trained and evaluated the latest architecture in stages, which improved the accuracy by about 10% in the short term (within 2 months) compared to the initial version.
- Partnered with data scientists to refine data inputs, embeddings, and augmentation strategies. Combination of superior models and data conditioning contributed to state-of-the-art results on key benchmarks.

SHANGHAI ANGDI ADVERTISING

Shanghai, China

Unity 3D Engineer

2020.9 - 2022.6

Engineered complex interactive programs by integrating Unity3D with diverse hardware sensors (Kinect, RealSense, Lidar). Delivered high-profile commercial projects, including real-world game recreations and biomechanical measurement tools for global brands.

Interactive Program

- Rebuild the game Pacman in the real world, using Unity3d to build the program and communicate with hardwares such as lights, signal transmitter and signal tower to locate Pacman and ghosts.
(Related video: <https://www.bilibili.com/video/BV1Mf4y1y7Zu/>)
- Use picture scanner and Unity3D to create a coloring game that after coloring the specific patterns which are provided by us in the paper, the picture will appear as a mesh in a specific 3d model which will be created immediately in the game scene.
- Collaborated with On Holding AG's foreigner team to make a place for customers to measure their feet size by using Unity3D with 4 RealSense cameras and an around 10 meters track that can calculate the runner's type by measuring the running speed with 2 Kinect.
- Developed an eye tracking application on the phone with real-time facial recognition for JINGDONG.

FREELANCER DIGITAL ART

Shanghai, China

Unity 3D Engineer, Blender Developer

2019.8 - 2020.7

AALab

Collaborated with AALab (digital art studio) on executing end-to-end development of AR mobile games and interactive installations using generative coding and 3D modeling. Collaborated with major platforms like BiliBili on CG production and led technical workshops on creative coding.

SHANGHAI YI ZUAN NETWORK TECHNOLOGY

Shanghai, China

Software Engineer

2018.10 - 2019.7

App game and H5 game

Architected scalable backend infrastructure and CI/CD pipelines on AWS and Tencent Cloud for high-traffic mobile games. Optimized production environments through load balancing, database clustering, and performance analytics.

SHANGHAI INTUQ INFORMATION SECURITY TECHNOLOGY

Shanghai, China

Software Engineer

2016.6 - 2018.9

Promoted to Team Lead within six months to head the development of a security vulnerability management SaaS platform. Built a distributed microservices architecture from the ground up, implementing automated scanning, data visualization, and task queue systems.

The security vulnerability management SaaS platform

- Developed and led projects as part of a 3-member team using Python, Django, Celery, MySQL, Redis, RabbitMQ, MongoDB, Angular
- Created scripts mimicking white hat techniques to gather asset data from RPC sources
- Constructed microservices architecture; implemented task queues and workers with Celery; Handled server ops, infrastructure management in distributed deployment
- Built data visualization reports incorporating pandas for analytics and charts

EDUCATION

XU EXPONENTIAL UNIVERSITY OF APPLIED SCIENCES

POTSDAM, GERMANY

M.SC.DATA Science

2025.10 - 2027.9

ARIZONA STATE UNIVERSITY

TEMPE, ARIZONA, USA

Master of Engineering, Specialized in Machine Learning

2023.9 - 2025.7

GPA: 4.0/4.0

CHENGDU UNIVERSITY OF INFORMATION TECHNOLOGY

CHENGDU, SICHUAN, CHINA

Computer Science and Technology (Application Technology)

2013.9 - 2017.6

PUBLICATIONS

LIU, B.; XU, Y. SPARSITY LIMIT TO PRUNE LARGE LANGUAGE MODELS FOR ON-DEVICE AI ASSISTANTS: LLAMA-2 AS AN EXAMPLE.

PREPRINTS 2024, 2024071568.

SKILLS

- **Languages:** Proficient in Python; Working knowledge of Golang, C#; SQL, NoSQL, Bigtable
- **Platforms:** Skilled in Linux (Ubuntu, CentOS); Proficient in Windows and macOS environments
- **Machine Learning & AI:** End-to-end ML model delivery, deployment & integration in cloud environments, n8n
- **Backend:** Redis, RabbitMQ, Celery for distributed systems, k8s
- **Cloud:** Google Cloud proficient; Familiar with AWS, Azure, Alibaba Cloud
- **3D Applications:** Blender for 3D modeling and animation; Houdini for effects and procedural modeling
- **Game Engine:** Unity3D expert with emphasis on
 - Visual Effect Graph for high-fidelity graphics
 - Physics and post-processing pipelines
 - Hardware integrations (AR/VR, sensors, control devices)
 - C# scripting and shader programming
- **Methodologies:** Agile, CI/CD, microservices, containerization

SELF ASSESSMENT & INTERESTS

- **Multi-industry Exposure** - My varied background spanning diverse sectors allows me to analyze problems through multiple lenses and devise creative solutions rapidly.
- **Quick Learner & Adaptable** - With strong self-study skills and adaptability to new environments, I efficiently pick up new languages, frameworks, and development tools. I learn by doing.

- **Disciplined & Motivated** - I set clear milestones and drive to exceed them. Skilled at managing my own projects and deliverables while collaborating cross-functionally.
- **Leadership & Ownership** - Experienced leading engineering teams and complex technology projects. I interface with stakeholders while driving engineering excellence and product success.
- **Communication** - Articulate technical concepts clearly to both business and engineering audiences. English CEFR C1.
- **Lifelong Learner** - Innately curious to increase my competencies. I read research papers and blogs about the latest ML innovations for continued growth.
- **Interests** - Freediving, hiking, baking, visual storytelling through experimental photography, crafting 3D animations in Blender