

Sebastian Ikin, Senior Software Engineer and AI Founder

Senior software engineer with five years in Arm's GPU organization and over two years as co-founder, CTO and CEO of an AI company. Seeking technically and strategically challenging roles in software engineering, AI infrastructure, platform development and technical leadership. Based in Oslo, bilingual Norwegian/English.

[GitHub](#) | [LinkedIn](#) | sebastianaikin@gmail.com

Experience

March 2024 - present: CEO / CTO / Co-founder, Rafmer AS

Co-founder and technical lead in a two-person AI company developing software and infrastructure for local AI inference and agentic workflows. Started as Chair and CTO, becoming full-time CEO in June 2025. Responsible for technical architecture, product direction and operations.

Technology and Product

- Designed and built an internal **Kubernetes-based GPU cluster** from hardware through the application layer for development, testing and operation of local language models.
- Built inference deployments on consumer GPUs using **llama.cpp, vLLM, CUDA and Kubernetes**, including five RTX 3090-based deployments supporting eight concurrent clients and approximately 50 tokens/s per deployment with Qwen3.8-27B.
- Optimized LLM inference with a focus on **GPU throughput, memory usage, context length, parallelism and cost per request**, without enterprise GPU virtualization.
- Designed and productized an **edge AI server based on AMD Ryzen AI Max**, Ubuntu, k3s, Vulkan and llama.cpp, including hardware selection, CAD design, prototyping and the software stack.
- Developed backend and infrastructure services primarily in **Python and FastAPI**, deployed with Kubernetes, Helm and YAML-based workloads.
- Designed and developed **REST APIs, MCP servers, API gateways, model routing and RBAC systems** for internal and customer-facing services.
- Built internal tools and agent integrations for **cluster observability, project management, web search and developer workflows**.
- Designed and deployed a **GitHub-integrated developer agent** running on the company's own AI infrastructure, capable of working with internal repositories and development processes.
- Designed and built customer-facing **ETL pipelines with Apache Airflow**, including customer-specific access control and processing of external datasets.
- Built and delivered customer-facing datasets through **external REST APIs** with isolated projects and access control.
- Developed a lightweight **OpenRouter gateway integrated with Supabase and Stripe** for subscription validation, API access and model routing.
- Tested and optimized **local AI inference on Linux and Android**, with particular focus on hardware compatibility, throughput, model quality and memory constraints.
- Contributed to the architecture and quality assurance of the company's **Flutter-based Agent Platform**, including the inference runtime, API integrations and local model execution.

Leadership and Business

- Responsible for the company's technical direction, product priorities and project management, primarily through GitHub Projects.
- Led a strategic pivot from **B2B sales of AI hardware and licensing to local B2C inference** on consumer devices with a SaaS business model.
- Completed and documented a **market validation project funded by Innovation Norway**.
- Worked with financial modelling, budgeting and cost management using Fiken and Google Sheets.
- Participated in investor and VC discussions, fundraising work and strategic assessments related to capital requirements and the company's continued development.
- Negotiated, drafted and coordinated agreements such as **NDAs, letters of intent and customer agreements**.

- Responsible for the company's work on **GDPR, information security and the EU AI Act**.
- Managed GPU and AI hardware procurement from Norwegian and international suppliers, focusing on performance per krone.
- Worked with market analysis, product pages, technical communication and digital marketing via Google Ads, LinkedIn and Meta.

June 2020 - May 2025: Senior Software Engineer - GPU, Arm Trondheim

Worked in Arm's GPU organization developing and maintaining Python tools for GPU test automation, GPU bring-up, data processing and analysis. Started part-time during my studies, later moving full-time. Progressed to Senior Software Engineer, responsible for key internal tools, data deliveries and cross-team support. Selected contributions:

- Built and maintained an **Apache Spark-based ETL pipeline** for automated issue detection in large test datasets.
Reduced manual filtering and saved the team about **eight working hours per week**.
- Built a research-based **image verification algorithm and data pipeline** automating a previously manual classification process.
Achieved about **70% precision** and cut manual work by about **five hours per week**.
- Built and maintained an internal **React/Node application** for data classification/analysis.
- Managed **Jenkins pipelines** for test, workflow and release automation.
- Improved testing and QA for internal data-distribution tools, helping reduce critical production errors to near zero over roughly three years.
- Troubleshoot complex issues across **hardware, drivers, test infrastructure and software** during GPU bring-up.
- Identified and escalated cross-team test issues and contributed to root-cause analysis and coordination.
- Delivered analyses and datasets to four internal teams, technical leads and external partners.
- Contributed code reviews, architecture discussions and improvements to internal development processes.
- Ran **technical interviews** and helped recruit developers.
- Mentored and onboarded new hires and supported less experienced developers.

Education

August 2018 – June 2021: Bachelor of Engineering - Computer Engineering, NTNU Trondheim
Specialization in software architecture. Bachelor's thesis in **penetration testing and cybersecurity**.

Technical Skills

Programming: Python, TypeScript, SQL, JavaScript, Java, Bash, C++, Dart / Flutter

Backend & Software Engineering: FastAPI, Node.js, REST APIs, microservices, MCP, RBAC, PostgreSQL, Supabase, serverless, distributed systems, system architecture, OOP, TDD, design patterns, code review

Data and ML: Apache Spark, Apache Airflow, Pandas, NumPy, PyTorch, ETL / pipelines, data quality, validation, GPU inference

AI Infrastructure / MLOps: Kubernetes, k3s, Helm, llama.cpp, vLLM, CUDA, ROCm, Vulkan, TensorRT-LLM, Triton Inference Server, Nvidia GPU, AMD Ryzen AI, Arm Mali, local LLM inference, model/GPU/memory optimization

DevOps and Observability: Git, GitHub, Gerrit, GitHub Projects, Jenkins, Rancher, Docker, Prometheus, Grafana, Loki, Linux / Ubuntu

Cloud: AWS EC2, Lambda, RDS, Aurora, VPC, S3, IAM, GCP Firebase, Firestore, Supabase PostgreSQL, Auth, Edge Functions, Storage, Vault, RLS, Cloudflare Gateway

Leadership & Strategy: Technical leadership/strategy, product development, project management, architecture, mentoring, onboarding, technical interviews, business development, financial modelling/cost management, market analysis, investor/stakeholder communication, GDPR, information security, EU AI Act, SCRUM, Kanban, Jira, Confluence, GitHub Projects