

MICHAEL VASILYEV

Senior Full Stack, Backend & AI Engineer | TypeScript | Node.js | React | Svelte | Java | Python | FastAPI | AWS | Serverless | Liferay | Computer Vision | LLMs

Netanya, Israel | 053-9354048 | vasilyevmichael78@gmail.com | linkedin.com/in/michael-vasilyev-5a8ba8203 | Languages: Hebrew, English, Russian

PROFESSIONAL SUMMARY

Senior Full Stack & AI Engineer with 9+ years of hands-on experience building production web platforms, distributed microservices, internal operations tools, backend services, AWS serverless workflows, and AI/computer-vision infrastructure. Strong recent focus on AI-powered vehicle inspection systems, FastAPI inference services, Computer Vision, LLM/VLM infrastructure, RAG/embeddings, self-hosted inference, Liferay portal development, and AI-assisted software delivery.

CORE TECHNOLOGIES

Languages	TypeScript, JavaScript, Java, Python, Bash
Frontend	React, Svelte 5, Vite, Material UI, Blueprint, React Bootstrap, Zustand, Redux, TanStack Query, Axios, CSS/Sass, responsive UI, admin panels
Backend	Node.js, Express.js, NestJS, AdonisJS, Java 11+, Spring Boot, Liferay modules, Python, FastAPI, REST APIs, serverless functions, microservices
AWS / Cloud	AWS Lambda, AWS SAM, ECS, EC2, S3, Cognito, SSM, Secrets Manager, IAM, Directory Service, CloudWatch, Docker
Data / Messaging	MySQL, PostgreSQL, MongoDB, Redis, DynamoDB, RabbitMQ, Kafka, NATS, Elasticsearch, etcd, LDAP/Active Directory integrations
AI / CV	FastAPI inference services, SAM/SAM3-style segmentation, YOLO11, DINOv3, MiniLM embeddings, Qwen VLM, vLLM, llama.cpp, Hugging Face Transformers, PyTorch, OpenCV, Pillow, LangChain, LangGraph, RAG, embeddings, vector databases, OpenAI/Gemini/Anthropic APIs
AI Coding	GitHub Copilot, OpenAI Codex, Cursor IDE, Claude Code, AI-assisted code generation, review and refactoring workflows

PROFESSIONAL EXPERIENCE

RAVIN.AI - Senior Full Stack & AI Engineer

Haifa, Israel | Dec 2021 - Present | AI-powered vehicle inspection platform

Role Overview

Builds and maintains internal software used to operate, configure, monitor, and extend Ravin AI's vehicle inspection platform. Work spans frontend admin products, Java/Liferay portal modules, Node.js services, AWS Lambda functions, media processing, production alerting, identity/invitation workflows, and Python/FastAPI inference services for vehicle-image understanding.

Key Contributions

- Delivered 4+ years of production features combining web applications, backend services, AWS cloud services, and Computer Vision workflows.
- Built and maintained Liferay-based internal portal tools for operations, support, configuration, and engineering teams, including tenant/site/user management, CRM/configuration flows, monitoring, Redis and etcd inspection, manual processing, external-image ingestion, invitation links, reports, service versions, and data trailers.
- Developed Liferay client extensions and supporting Java/Liferay APIs using React, Svelte, Vite, Material UI, Blueprint, React Bootstrap, Zustand, TanStack Query, Axios, Java, Gradle, and Liferay workspace tooling.
- Developed and maintained TypeScript/Node.js microservices and backend services supporting production microservice architecture, internal platform capabilities, media workflows, and operational tooling.
- Built and maintained AWS Lambda/serverless workflows for invitation creation, expired invitation cleanup, temporary Cognito user cleanup, inactive IDP user cleanup, IDP login handling, compressed media/event ingestion, tenant/Cognito pool resolution, AWS Directory/Active Directory integration, Active Directory federation, Cognito federation, SSO/IDP login workflows, S3 ZIP processing, media classification, and event/media metadata creation.
- Built media-handling backend services in NestJS/TypeScript for S3-backed media workflows, RabbitMQ integration, scanner integration, anonymization/video/simple-image flows, downstream alerting, Docker-based deployment, and request/config management.
- Implemented Slack-integrated operational alert service with Express.js, alert payload validation, API rate limiting, RabbitMQ queue/consumer checks, CloudWatch Logs queries, daily metrics tracking, Redis/local storage support, and health/status endpoints.
- Built Python/FastAPI inference infrastructure for vehicle-image workflows, including sync HTTP routes, async RabbitMQ processing, central configuration, S3 model/config bootstrap, request tracing, cache management, alerts, health/config endpoints, and multi-plugin model loading.
- Worked on Computer Vision pipelines using SAM/SAM3-style vehicle-part segmentation, YOLO11 beauty-shot detection/classification, DINOv3 image embeddings, MiniLM embeddings/search, Qwen VLM support, triage/rerank workflows, result/image/green-background caching, S3 path handling, and self-trained/custom YOLO model workflows.
- Built local/self-hosted LLM and VLM runtime tooling using Docker, vLLM OpenAI-compatible serving, Qwen-VL/Qwen model configuration, S3 model synchronization, llama.cpp local serving, and browser-based vLLM metrics monitoring.
- Reduced dependency on external OpenAI and Claude API calls for selected workloads by using self-hosted vLLM/VLM inference and local model-serving workflows.
- Designed AI/RAG experiments and services using LangChain, LangGraph, embeddings, vector databases, MiniLM, OpenAI API, Gemini API, and Anthropic API for retrieval-based AI workflows, semantic search, LLM orchestration, and AI-assisted application features.
- Worked across TypeScript, Node.js, Express, NestJS, Python, FastAPI, Redis, RabbitMQ, AWS ECS/Cognito/SSM/Secrets Manager/EC2/CloudWatch/S3/IAM/Directory Service, serverless functions, Java 11+, React, Svelte, Liferay Portal, Active Directory, MySQL, and Docker environments.

MediWho - Full Stack Developer

Haifa, Israel | 2021 - Dec 2021 | Healthcare / insurance platform

- Developed healthcare insurance web platform and CRM/admin functionality for managing policies, privileges, and application workflows.
- Implemented payment-related features and customer-facing flows.
- Worked across frontend and backend application development.

Tech stack: React, Next.js, Node.js, NestJS, AdonisJS, TypeScript, Elasticsearch, Docker.

Elpisor LTD - Full Stack Developer

Rehovot, Israel | 2017 - 2021 | E-commerce, ticketing, and healthcare systems

- Built e-commerce and delivery-management platform features including user roles, order management, worker management, search, and statistics.
- Developed event-driven ticketing platform components and shared backend libraries.
- Built backend services for healthcare sensor-monitoring workflows.

Tech stack: JavaScript/Node.js, Java, Spring Boot, NATS, Kafka, REST APIs, MongoDB.

EDUCATION

Master's Degree in Mathematics and Computer Science

Ural State University, Yekaterinburg, Russia | 2002 - 2007

TARGET ROLES

Senior Full Stack Engineer | Senior Software Engineer | Senior Backend Engineer | AI Engineer | AI Platform Engineer | Cloud Backend Engineer