VADIM POLIAKOV

Senior AI/ML Engineer
vadimpoliakov99@gmail.com / linkedin.com/in/vadim-poliakov-ai / vadimpoliakov.dev

SUMMARY

AI/ML Engineer with 7+ years of experience (5+ years commercial), T-shaped across FullStack, BackEnd, FrontEnd, QA, and DevOps. As the field of AI has been revolutionizing throughout the decade, I have built and scaled several AI products from idea to production that improved key business metrics, accelerated scaling into new markets, and drove stakeholder interest in investment.

KEYWORDS

Python; AI; ML; AI/ML; LLM; PostgreSQL; AWS; LangChain; AI Agents; RAG; CI/CD; TypeScript; FullStack; Docker; LangGraph; React; Next.js; Node.js; Kubernetes; BackEnd; Cloud; Azure; Data; workflow; evaluation; Claude; API; REST API; FastAPI; architecture; harness; Vector DBs; Databases; PGVector; retrieval; semantic; Vue; Redis; FrontEnd; JavaScript; Git; GCP; SQL; NoSQL; MongoDB; OpenAI; Kafka; PyTorch; VLM; multimodal; Django; Terraform; Cursor; Angular; LlamaIndex; RabbitMQ; Anthropic; LoRA; QLoRA; fine-tuning; MySQL; CSS; Gemini; NLP; Conversational AI; tool calling; chunking; Pinecone; S3; DynamoDB; GitHub Copilot; DevOps; LLMOps; Prompt; Prompt Engineering; Linux; Microservices; MCP; HTML; UI/UX; Tailwind CSS; Grafana; Prometheus; Flask; NestJS; Embeddings; scikit-learn; Hugging Face; Computer Vision;

WORK EXPERIENCE

MIEA PTE. LTD. | Senior AI/ML Engineer Feb 2025 – Present

A healthcare agentic AI assistant that automates front-desk operations by communicating with patients through messaging platforms (WhatsApp and others). It handles scheduling, routine questions, appointment reminders, follow-ups, and feedback collection. It is integrated with 20+ clinic management services.

- Revived this project by rewriting the legacy solution and architecture into a new modular one with frameworks for AI (Python, TypeScript, architecture, system analysis, LangGraph, LangChain, JavaScript).
- Made clinic documents usable for the AI by ingesting them with LlamaIndex into PostgreSQL plus PGVector (LlamaIndex, PostgreSQL, PGVector, Vector DBs, RAG, Python, BackEnd, SQL, ETL).
- Made an AI Agent with RAG in production by writing agentic pipelines and implementing tool calling with LangChain and LangGraph (LangChain, LangGraph, AI Agents, tool calling, RAG, workflow, Python).
- Added observability with LangSmith on retrieval functions and reranking functions so each turn showed what was retrieved, what was reranked, and what the agent answered (LangSmith, retrieval, reranking, LLMOps, LangGraph, LangChain).
- Increased the quality of retrieval by using advanced techniques (multi-query, hierarchical retrieval, reranking, BM25, RRF, different chunking strategies, structure-aware chunking, MMR).
- Kept clinic and patient data fully private by running the local LLM of the AI agent self-hosted on a cloud server. This way the data was not exposed to ChatGPT (OpenAI), Claude (Anthropic) and Gemini (Google). (Cloud, AWS, BackEnd, Azure, GCP, private LLM, self-hosted, data privacy).
- Solved critical prompt-injection vulnerabilities and credential exposure risks by adding a prompt guard and implementing MCP as a secure middleware layer for calling internal microservices and third-party clinic management software (MCP, Prompt, architecture, API, Microservices, FastAPI, FastMCP).
- Scaled adoption among clinic clients across multiple platforms by connecting AI to messaging apps (WhatsApp, Viber and others) (BackEnd, webhooks, Python).
- Kept the AI agents online while we shipped new features by implementing a microservice architecture and deployments with Kubernetes. Rolling deploys ran those same services in production with no clinic downtime. The testing environment was in Docker, Docker Compose (Kubernetes, Docker, Microservices, Pods, Docker Compose, QA, testing).
- Increased the speed of RAG by using Redis for caching frequently asked questions across multiple users (Redis, BackEnd, embeddings).
- Kept the AI chat with customers responsive at all times by executing agentic actions asynchronously via RabbitMQ (event-driven architecture, task queuing, async, execution, async messaging).
- Improved the UI/UX of the web app for clinic managers by showing how the AI thinks and where it retrieved its information (FullStack, FrontEnd, Next.js, BackEnd, React, React.js, JavaScript, Python, UI/UX, SSE, React Hooks).

INBECA WELLNESS EQUIPMENT S.L. | **Senior AI/ML Engineer** Dec 2023 – Feb 2025

An AI Agent that helps call-center employees and field technicians find information in the company's manuals, product documentation, schematics and repair guides. It gives quick access to the company's technical knowledge. It helps support staff handle customer enquiries and provide initial troubleshooting advice, while accelerating the onboarding of new technicians who need to learn repair and maintenance procedures.

- Made it possible to start the project by turning the business idea into a tech plan. Dived deep into the company's support processes, field-technician processes, and company data.
- Explained common problems with AI in production systems, showed potential pitfalls, negotiated where we could and should cut costs and/or time, calculated risks, and proposed metrics to evaluate the AI agent. Established the tech stack for the AI system.
- Made company data usable for the AI by putting unstructured enterprise documents (call-center history, documentation, PDFs) into a DataLake in Cloud using S3 and MongoDB, embeddings. (Cloud, AWS, Azure, GCP, PGVector, Vector DBs, Data, Databases, NoSQL, Python, BackEnd, Embeddings, MongoDB, S3).
- Validated that the idea of the RAG Agent on said DataLake was feasible by making an AI agent with RAG. Implemented prompt engineering, retrieval by semantic search, keyword search, BM25, and citations to ground the AI in the company's data and reduce hallucinations (Python).
- Greatly accelerated development by creating validation and test data samples for checking the correctness of AI responses. This way, I did not halt development for human review every time I made a new version of the AI.
- Gave the resulting AI to the call-center staff and field technicians by making the FrontEnd web app. Collaborated with the FrontEnd developer and used FastAPI, Docker, Docker Compose to run the BackEnd on the company's cloud server (FrontEnd, Next.js, Cloud, AWS, BackEnd, API, FastAPI, Docker, Docker Compose, React.js).

SINGULAR SOFT GROUP PTE. LTD. (prev. CSSSR PTE. LTD.) | **AI/ML Engineer** Sep 2021 – Dec 2023

Heart CV (iOS app): AI-powered health tracker. It enables patients to monitor blood pressure and glucose levels on their personal devices by simply scanning medical devices with their phone camera.

- Tripled user conversion and retention by building a Computer Vision pipeline (PyTorch, YOLO, OpenCV, Python, LSTM, OCR) integrated with iOS (CoreML, ONNX, Swift, C++). Built, trained, validated, and tested a custom deep learning model from scratch. Further optimized the pipeline for real-time on-device inference.
- Cut development cost and time by making a data augmentation framework simulating real-world conditions that users might have problems with (different fonts, blur, glare, skew, poor lighting, rotations).
- Opened the Singapore medical market for the company by developing and deploying new features in glucose recognition and increasing the quality and execution speed of existing AI/ML algorithms.

Space Research Institute | **ML Engineer** Jan 2019 – Mar 2022

- Made advancements in the retrieval of total precipitable water in the atmosphere over land, using radiothermal monitoring data from multiple satellites. Wrote several scientific articles and gave a conference talk.
- Designed and trained custom neural networks in PyTorch, Keras, TensorFlow.
- Implemented several machine learning algorithms in projects for analysis of the equatorial Pacific thermocline.

EDUCATION

B.Sc. Computer Science and Engineering

Moscow Technical University of Communications and Informatics

Thesis: Design and Development of a Candidate Selection System for the IT Market. A local Gradio app. Saiga Mistral 7B was fine-tuned (LLM, SFT, PEFT, LoRA, QLoRA) on IT resumes with custom evaluation and testing metrics.

LANGUAGES

English C2 **Japanese** B1