# Alexey Zaytsev

AI/ML Engineer & Systems Architect

Full-stack ML engineer specializing in generative AI, distributed inference, and model fine-tuning. Combines first-principles deep learning with low-level systems and hardware engineering.

https://alexey.work/cv
alexey.zaytsev@gmail.com
https://github.com/xl0
https://www.linkedin.com/in/alexeyzaytsev

## // FEATURED PROJECTS

**lovely-tensors** 1.4k ★ / **numpy** / **jax** — Python, PyTorch, NumPy, JAX
Tensors for human consumption. Widely adopted by AI researchers.

**tidygrad** — Python, NumPy, Autograd
From-scratch autograd engine, enough to train a tiny GPT-2 style model.

**congusto-chat** — SvelteKit, TypeScript, PostgreSQL, Vercel AI SDK
Open-source multi-provider chat interface, similar to OpenWebUI and t3chat.

**lovely-docs** — SvelteKit, TypeScript, Python, LLM, MCP
AI-native documentation dehydrator for agentic access. CLI, web UI, and MCP server.

**pelican** — SvelteKit, TypeScript, PostgreSQL, Vercel AI SDK
LLM-driven SVG/ASCII art generation gallery with iterative visual refinement.

**latent-tools** — Python, PyTorch, ComfyUI, Diffusion Models
Custom ComfyUI nodes for fine-grained diffusion latent manipulation.

**nvml-tool** — C, NVML, CUDA, Linux
GPU power and fan control utility built on NVIDIA NVML.

## // EXPERIENCE (ML SYSTEMS)

### AI/ML Engineer (Consulting)

July 2024 - Present

Select client engagements:
• AI Storyboarding Platform (Python backend): Integrated new image generation providers and refactored the generation abstraction layer across use cases (assets, scenes, inpainting).
• Earnings Call Analysis & CEO Training: Engineered a low-latency transcription + LLM pipeline (AWS Transcribe / Deepgram) with speaker diarization and financial metrics extraction. Built a companion CEO training module that simulates analyst Q&A via TTS, transcribes spoken responses, and scores both content quality and delivery (WPM, filler words).
• Open-source AI projects: congusto-chat, lovely-docs, pelican, latent-tools, nvml-tool.

**Python · LLM Pipelines · AWS Transcribe · Deepgram · Speaker Diarization · TTS · Latency Optimization · Applied AI**

### Generative AI Lead

Nov 2023 - July 2024

io.net

Led development of the generative AI platform supporting image, video, and 3D model generation. Served tens of thousands of users, generating hundreds of thousands of images:
• Architecture & Inference: Built an end-to-end generation pipeline. The frontend talked to the backend that talked to Supabase/PostgreSQL, which queued tasks via RabbitMQ. Distributed GPU workers picked up the jobs, generated the media, uploaded to S3 (served via CloudFront+imgproxy), and pinged a backend webhook, which finally streamed updates to the client via Server-Sent Events.

## // RELEVANT SKILLS

**Core ML**
Python, PyTorch, JAX, CUDA, Transformers, Diffusion Models, Computer Vision, Autograd, CLIP, Embeddings, Weights & Biases

**Applied AI**
LLM Agents, vLLM, Hugging Face, LangChain, RAG, Fine-tuning, LoRA, Stable Diffusion, Vector Databases, Vercel AI SDK, MCP

**Full-Stack & Platform**
SvelteKit, TypeScript, FastAPI, PostgreSQL, Tailwind CSS, SQLAlchemy, Drizzle ORM, Docker, Linux, AWS, GitHub Actions, Supabase, Vercel, CloudFlare, S3

## // EDUCATION

**Université Denis Diderot**
2013 - 2015
M.Sc. Systems and Synthetic Biology

**LETI University**
2003 - 2009
Specialist, Computer Security

## // LANGUAGES

| | |
|---|---|
| English | Professional |
| Russian | Native |
| Ukrainian | Elementary |

• Model Fine-tuning: Extensive work with Stable Diffusion inference, custom model fine-tuning, and LoRAs to optimize generation quality and capabilities.
• Custom NSFW Detector: Trained and deployed an EfficientNet classifier to replace off-the-shelf filters with high false-positive rates, reaching production-grade precision directly in the distributed GPU worker pipeline.
• MLOps & DevOps: Orchestrated and scaled the distributed inference infrastructure.

**Stable Diffusion · PyTorch · RabbitMQ · PostgreSQL · Supabase · LoRA · EfficientNet · MLOps · SSE · S3**

### AI/ML Engineer (Consulting)

Nov 2022 - Nov 2023

Independent consulting focused on AI applications and prototypes. Select engagements:
• Dog Finder: An experimental missing-dog app exploring both facial recognition and noseprint uniqueness. Trained EfficientNet with triplet loss.
• SAT Practice Generator: A dynamically updating test prep system seeded by real data. Built with Streamlit, Airtable, and OpenAI LLMs-migrating from GPT-3.5 to GPT-4 largely solved our initial correctness and consistency struggles.
• GPTEditor: An AI-assisted copywriting tool-like Cursor, but for prose. Users highlight a paragraph and instruct the LLM on how to rewrite it. Built with SvelteKit, Tiptap, and Supabase.

**SvelteKit · OpenAI API · RAG · Supabase · Streamlit · TypeScript · Vector DBs · PyTorch · Computer Vision**

### Electronics Engineering & Hardware Leadership

Sept 2015 - May 2020

Koi.science, Okra Solar, and Consulting

Led hardware product development from prototype to manufacturing across power electronics and embedded systems, while managing teams and production workflows in Shenzhen.

**Electronics Engineering · Embedded Systems · Power Electronics · Manufacturing**

### Research Engineering (Biology & Instrumentation)

Jan 2014 - Sept 2015

Institut Curie, Institut Pasteur, and CRI Paris

Built experimental tools and conducted wet-lab research in microfluidics and molecular biology, contributing to published scientific work.

**Bioinformatics · Molecular Biology · Microfluidics · Research**

### Low-level Systems Software Engineer

Jan 2005 - July 2012

AltEll, Protei, and Nexenta Systems

Developed kernel-level software (drivers, bootloaders, BSPs) for telecom and storage systems across ARM, MIPS, and x86 platforms.

**Linux Kernel · C · Device Drivers · Systems Programming**