Valeriy Selitskiy

Senior AI Infrastructure Engineer / Backend Lead

Production LLM and diffusion systems, local inference, quantization, privacy-aware AI products, and high-throughput Go/Python platforms.

+48 571 203 032
wavecutdnb@gmail.com
linkedin.com/in/wavecut
github.com/iamwavecut
huggingface.co/WaveCut
t.me/WaveCut
Krakow, Poland (EU) - CET
Fully remote - B2B or employment

PROFILE

Senior engineer with 20+ years in backend and platform systems and a current focus on production AI infrastructure. I build the layer around models: serving, provider routing, fallback behavior, queues, observability, memory, privacy filtering, safety, cost control, and GPU-constrained deployment. Recent work combines startup backend leadership with hands-on local LLM and diffusion model operations, including quantization and inference optimization for real users rather than demos.

13k MAU / 300k LLM reqs per month

Built and operate Plotva, a local-first Telegram AI system with multimodal tools, memory, safety, analytics, and GPU infrastructure.

>99% local conversational path

Migrated normal dialogue traffic to self-hosted inference with RTX 3090 + RTX 4060 Ti services and external fallback for overload.

60 public Hugging Face model repos

Published quantization and local-inference artifacts across Diffusers, Transformers, MLX, GGUF/TurboQuant, EXL2, and bitsandbytes/AWQ formats.

Large models into runnable artifacts

Repeatedly convert oversized LLM and diffusion checkpoints into documented local-inference packages with measured memory, speed, and quality trade-offs.

CORE SKILLS

AI / ML Infrastructure
local inference, OpenAI-compatible endpoints, provider routing, prompt caching, batching, fallback paths, GPU queues, observability

LLM Runtime
PyTorch, Transformers, vLLM, llama.cpp, ExLlama/EXL2, GGUF, tool calling, agentic workflows, context management

Quantization
SDNQ, MLX, GGUF/TurboQuant, EXL2, AQLM, bitsandbytes, AWQ, quanto, ModelOpt, DWQ/GPTQ, FP8, NVFP4, UINT4/INT8, SVD

Diffusion / Multimodal
Diffusers, SD/SDXL, Flux, Sana, Lens, HiDream, image generation/editing, OCR, vision, music generation

Backend / Platform
Go, Python, PHP, TypeScript, PostgreSQL, pgvector, Redis/Dragonfly, Docker, Kubernetes, AWS, GCP, bare metal

Delivery / Operations
GitHub Actions, GitLab, Jenkins, ArgoCD, blue-green deployment, Prometheus, VictoriaLogs, Grafana, incident debugging

EXPERIENCE

Mobiways - Founding Engineer / Backend Lead

Jan 2025 - Present

Go, Python, Docker, GitHub Actions, bare-metal infrastructure, Prometheus, VictoriaLogs, Grafana, LLM and diffusion provider integrations.

- Own backend, platform, infrastructure, deployment, and internal tooling for an AI mobile-app startup as the end-to-end technical owner across architecture and operations.
- Design and maintain on-prem, bare-metal infrastructure: network topology, core services, deployment pipelines, release process, and operational controls.
- Build high-throughput Go services for heavy parallel workloads and unified internal APIs that integrate multiple LLM and diffusion providers.
- Automate CI/CD with GitHub Actions, Docker, and blue-green deployments, supporting zero-downtime releases and fast rollbacks.
- Implement production observability with Prometheus, VictoriaLogs, and Grafana; write infrastructure tools, data pipelines, and operational automation in Go and Python.
- Shape engineering roadmap and practices in a small startup environment where architecture, implementation, deployment, and incident response are tightly coupled.

AI SYSTEMS AND QUANTIZATION WORK

Plotva - Independent Founder / AI Systems Engineer

2011 - Present

Telegram AI product, local LLM inference, diffusion model serving, OpenAI-compatible APIs, pgvector memory, privacy filtering, Shield safety retrieval, GPU queues.

- Built and operate a non-commercial local-first social AI system with around 13,000 monthly active users and roughly 300,000 LLM requests per month.
- Designed a multi-user Telegram AI character system with dialogue, image generation/editing, OCR, translation, summarization, music generation, web search, analytics, long-term memory, and safety behavior.
- Migrated more than 99% of normal conversational LLM requests to self-hosted inference; production topology uses an RTX 3090 24GB for the main LLM workload and an RTX 4060 Ti 16GB for supporting AI services.
- Run local inference through a llama.cpp-based OpenAI-compatible endpoint with batching, large context handling, provider fallback for overload, and cost-aware routing.
- Built an AI farm around separate queues and execution paths for dialogue, image generation, image editing, memory processing, and supporting services, with backpressure, priority handling, and progressive user limits.
- Implemented long-term memory as an asynchronous background pipeline: chat-window processing, noise filtering, token estimation, extraction, PostgreSQL + pgvector storage, and hybrid retrieval with user/group boundaries.
- Added privacy and safety infrastructure: local PII redaction before model execution, privacy-aware fallback design, and Shield retrieval for high-risk topics such as self-harm, violence, doxxing, stalking, and dangerous instructions.
- Optimized around practical AI product constraints: VRAM limits, quantization artifacts, queue pressure, context size, prompt caching, fallback behavior, and model/runtime choice for each GPU.

Hugging Face Portfolio - Quantization, Local Inference And Model Packaging

2023 - Present

60 public model repositories as of Jun 2026 - MLX, Diffusers, Transformers, GGUF/TurboQuant, EXL2, SDNQ, ModelOpt, Quanto, FP8/NVFP4, bitsandbytes/AWQ.

- Maintain a public Hugging Face portfolio spanning text generation, text-to-image, image-text-to-image, image-text-to-text, music LM components, and encoder/model-component quantization.
- Published systematic SDNQ portfolios for HiDream O1 Image / Dev, Anima Preview 3, Lens/Lens-Turbo, ERNIE-Image, SDXS, Flux-family models, and other large text-to-image systems, covering static/dynamic 4-bit, UINT4/INT8, SVD, component-only, and full Diffusers packaging.
- Package local LLM artifacts for Apple Silicon and local runtimes: QClaw/OpenClaw MLX with DWQ/GPTQ variants, Qwen/DeepSeek/reasoning MLX models, ACE-Step music LMs, DeepSWE coding models, Gemma/Gemma4, Russian/roleplay EXL2 models, and bitsandbytes/AWQ conversions.
- Use Hugging Face model cards as reproducible engineering records: loading recipes, base-model links, tags, benchmark notes, generation examples, contact sheets, and explicit quality/speed/VRAM trade-offs.
- Turn research checkpoints that are expensive, awkward, or memory-heavy into practical local-inference artifacts, documenting what became smaller, what became runnable, and what trade-offs changed.

BACKEND TRACK RECORD

Lyft Oct 2019 - Dec 2021	Backend Software Engineer in Mapping/Search. Built and maintained search services, operations tooling, and public POI geospatial APIs; worked on data ingestion and ETL optimization, with combined cost-reduction efforts expected to exceed $10M; mentored engineers moving to Go and contributed to specs, architecture, and design reviews.
Unity Mar 2023 - Oct 2023	Senior Golang Engineer on a B2B contract under NDA. Optimized Go services, with old CV metrics recording 25% latency reduction and 30% reliability improvement; focused on refactoring, service design, error handling, and production readiness.
Independent Contractor Sep 2022 - Present	Develop and maintain web applications and backend services across Go, Python, and PHP; integrate third-party systems and keep services performant and reliable across varied teams and domains.
SoftSwiss Jul 2019 - Oct 2019	Designed and built real-time low-latency Go services for sports betting; worked with bare-metal Kubernetes, NATS, ZeroMQ, RabbitMQ, PostgreSQL, and third-party supply/demand integrations.
VIV Dev Nov 2018 - May 2019	Developed backend architecture for a mobile longevity coaching app; built services, testing protocols, deployment flow, CI/CD, and monitoring foundations using Go, Python, Node.js, AWS, and Docker/ECS.
Rakuten Viber Nov 2013 - Aug 2018	Senior PHP Developer for public and private APIs behind Viber Out, Sticker Market, Viber Wallet, and related services; production LEMP stack with Symfony, MySQL, Memcached, GitLab, and Jenkins.
Earlier Roles 2005 - 2013	Progressed from web developer to senior developer and team lead roles across Alutech, Admirals, Stylesoft, Seobility, and Minsk City Executive Committee IT Center, working with PHP, JavaScript, C#, SQL, financial systems, and public-sector web systems.

OPEN SOURCE AND PUBLIC ARTIFACTS

Selected GitHub Work

- Public GitHub profile with Go, Rust, C++, PHP, JavaScript, and CSS repositories dating back to 2010.
- Maintains practical clients and tools such as unofficial OpenRouter, Tavily, Serper, and MCP-related Go projects.
- Experimental llama.cpp forks explore TurboQuant, Gemma-family runtime work, and GPU-oriented KV/cache trade-offs.

Languages And Work Style

- Languages: English, Belarusian, Russian.
- Preference: fully remote roles, B2B contract or employment contract, EU/CET-friendly collaboration.
- Strengths: ownership, architecture, incident debugging, written technical communication, mentoring, code review, pragmatic delivery.

Target roles: Senior / Staff AI Infrastructure Engineer, ML Systems Engineer, Applied AI Engineer, Backend Lead for AI products, LLM/Diffusion Platform Engineer.