Avi Fenesh

AI Systems Engineer | LLM Inference, Multimodal AI, Model Training

Tel Aviv, Israel | +972-58-637-5837 | aviarchi1994@gmail.com | avifenesh.ai

GitHub | LinkedIn | Hugging Face

Summary

Software engineer with 5+ years in production systems, including 4+ years on AWS ElastiCache. Builds LLM and multimodal inference systems in Python, PyTorch, Rust and CUDA, from model selection and fine-tuning through evaluation, quantized serving and deployment. Maintainer of Valkey GLIDE; built the LLM-based reviewer used on 600+ Valkey pull requests.

Selected AI systems and research

- **memra:** Built a Rust and CUDA inference engine for Blackwell GPUs with OpenAI-compatible serving, NVFP4 quantization and text, image and video input paths. MTP speculative decoding reaches up to 2.3x llama.cpp throughput while correctness gates keep output token-identical to plain decoding.
- **Beta Gate:** Trained LoRA-gated Qwen3.5 models with GRPO. The 4B model raised MATH-500 anytime accuracy from 78.6% to 85.2% while cutting mean reasoning length from 4,223 to 2,202 tokens; gains held on 130 unseen problems. Released training code, 298,986 generation traces and automated claim checks.
- **Layby:** Built an end-to-end LLM serving workflow with a 421M-parameter return-time predictor, Python sidecar and vLLM/SGLang adapters. Cut p95 resumed-turn latency 41% versus an LRU CPU tier in slow-disk Qwen3-8B tests. Released ReturnBench, built from public agent and chat traces.

Experience

AI Systems Engineer | Independent research and open source *2026 - Present | Israel*

- Fine-tune, quantize and evaluate open models; publish NVFP4 builds, MTP draft heads in GGUF and trained adapters on Hugging Face.
- Contribute LLM fixes and features to vLLM, SGLang, LMCache and llama.cpp.
- Built agent-sh: 25 public repositories with 2,200+ GitHub stars, covering local model routing, deterministic CLI selection and reusable agent infrastructure across 50+ coding-agent platforms.

Software Development Engineer | Amazon Web Services, ElastiCache *Sep 2022 - Present | Tel Aviv*

- Build performance-critical components of a large-scale in-memory data store in Rust, Java, Node.js and C, including module persistence, crash recovery and cluster replication.
- Design zero-copy cross-language interfaces using FFI, Unix domain sockets and shared memory to connect high-throughput Rust cores to client languages.

Maintainer | **Valkey GLIDE** and **valkey-skills** *May 2023 - Present | Remote, open source*

- Maintain the official multi-language Valkey client, with a shared Rust core and Java, Python, Node.js and Go bindings; 120+ merged contributions to the main repository.
- Built and maintain Valkey's LLM-based PR reviewer, used on 600+ pull requests with repository checkouts and targeted builds and tests. Created valkey-skills for 50+ AI coding agents.

Earlier experience

Software Developer | Trafficpoint | Oct 2021 - Sep 2022

Officer, Team Commander | IDF special units (Maglan, Rimon) | Jul 2014 - Nov 2018

Skills

AI/ML: LLM integration, evaluation and deployment; multimodal inference (text, image, video); PyTorch, LoRA, GRPO, quantization, speculative decoding, CUDA, ONNX, vLLM, SGLang, llama.cpp.

Engineering: Python, Rust, C, TypeScript/Node.js, Java, Go; AWS, Docker, Linux, distributed systems, performance optimization, FFI, Valkey/Redis.

Details

Computer Science studies | The Open University of Israel | 2019 - 2022

Languages: Hebrew (native), English (professional)