Erkin Mukhamedov

Milpitas, CA | mukhamedov.erk@gmail.com | (334) 559-7855 | linkedin.com/in/erkin-mukhamedov | github.com/ErikLancer

Education

- M.S. in Computer Science from Sofia University, Palo Alto, CA (GPA 4.0) — Status - Enrolled
- B.S. in Business Administration, Marketing from Auburn University (GPA 3.95) — Status - Graduated

Work History

AI Creative Engineer at Call2Action — *December 2025 to Present*

- Designed and built an enterprise-grade internal AI platform with a custom agent harness for domain-specific marketing workflows, including creative generation, landing page generation and hosting, motion editing, lead sourcing, and lead pipeline optimization, supporting marketing for 30+ enterprise clients concurrently.
- Developed the harness that runs the agent loop: delivering context and tools to the model, executing tool calls and handling their failures, streaming results with low latency, and managing session memory with checkpoint and restore in an isolated sandbox per session.
- Built a pluggable tool-use layer with function calling, MCP integrations, and structured output, connecting video, image, and voice generation models (Seedance, Kling, ElevenLabs) and business systems to the agent.
- Designed end-to-end agentic workflows from intent recognition to action execution and feedback, using a classifier router for skill selection, background tasks, and a director-reviewer critique loop for video editing.
- Secured the platform with layered defenses against prompt injection and excessive agency: classifier-model screening of untrusted input, separation of untrusted content from instructions, least-privilege tools, credential isolation so agents never hold secrets, sanitized model-generated output, and human approval for paid or irreversible actions.
- Built evals to diagnose and improve agent behavior using A/B tests, LLM-as-judge scoring, and tracing, making routine lookups 2x faster and 3x cheaper with no loss in quality.
- Migrated production to serverless containers with minimal downtime and optimized agent response time, startup, and compute cost using prewarmed sessions, shared caching, and automatic sleep of idle sessions.
- Analyzed production render history and $122k of Meta ad spend across 14,330 leads to identify winning creative patterns, and built prompt validators that prevent costly re-renders.
- Before the platform, developed individual AI agents for team members on the open-source Hermes framework (ad creatives, landing pages, social content, finance reporting, lead research), delivering 1,508 video ads for 53 clients.
- Built lead infrastructure on Cloudflare Workers: landing pages that deliver leads to the CRM, Meta Conversions API, and TikTok Events, sourcing pipelines that classify leads from Facebook groups and Craigslist, and a Stripe-to-Meta budget ledger with read-only agent access and audited writes.

Projects

Fine-tuned LLM for video ad prompt composition — *2026*

- Set up the evaluation pipeline first: a held-out client split, task metrics based on production validators, and blind LLM-judge comparisons against the base model.
- Built the training dataset from production prompt and render history labeled by outcome, supplemented with public Hugging Face datasets and synthetic examples generated by larger LLMs.
- Fine-tuned Qwen3.8-27B in Unsloth on cloud GPUs with both full-precision LoRA and 4-bit QLoRA, comparing output quality and memory cost, while monitoring evaluation loss and tuning the learning rate for stable convergence.
- Served the model in the cloud with vLLM and locally with Ollama using a quantized GGUF build, comparing latency and output quality across both.

paperbot - AI procurement bidding platform (side project) — *September 2026 to Present*

- Built a platform on Cloudflare that monitors government tender platforms around the clock, classifies and scores each lot on fit and margin, and prepares bids with an AI agent whose tools are generated from a typed API.
- Protected against prompt injection by separating untrusted tender text from structured fields and gating every outward action through a policy engine, a fail-closed classifier model, and human approval, with an isolated signing service.

Founder at SAKE - AI support for international students — *September 2024 to July 2025*

- Led development of an AI platform that helps universities support international students, pitched investors, and won admission to a leading Alabama startup accelerator.

Skills

- **Languages:** Python, TypeScript/JavaScript, Rust, SQL, Bash
- **Agents and LLMs:** agent harnesses (Pi, Hermes), LangChain, LangGraph, Vercel AI SDK, AutoGen, CrewAI, tool use and function calling, MCP, multi-agent orchestration, agent memory, RAG, vector databases, context engineering, prompt engineering, evals (A/B, LLM-as-judge), observability (Langfuse, OpenTelemetry)
- **AI security:** prompt injection defense, guardrails, input and output validation, least-privilege tool design, agent sandboxing, secrets isolation, human-in-the-loop approvals, OWASP Top 10 for LLM Applications
- **ML:** Unsloth, Hugging Face (Transformers, Datasets, PEFT), PyTorch, LoRA/QLoRA fine-tuning, evaluation pipelines, synthetic data generation, quantization (GGUF), inference with Ollama and vLLM
- **Cloud and infrastructure:** Cloudflare, AWS, GCP, Docker, Kubernetes, CI/CD (GitHub Actions), Linux