Alexei Korol: AI/ML Engineer

Greater Seattle Area | korolalexei@gmail.com | github.com/alexkorol

Profile

AI/ML engineer building and evaluating LLM-powered systems: retrieval, agent context tools, judge pipelines, local inference, and generative workflows. Focused on measurable behavior, explicit failure modes, and tools that remain inspectable after the demo.

Selected engineering work

Repo2GPT: Agent context infrastructure

Built a Python CLI, FastAPI job service, React control panel, and MCP server that produce language-aware repository maps and token-bounded code artifacts for coding agents. Added persisted jobs, SSE progress, authentication, health checks, and test coverage for the processing and MCP paths.

SongCraft RAG: Evaluated retrieval system

Built a 45-document, 2,376-page RAG corpus with local BGE-small embeddings, BM25+dense fusion, cross-encoder reranking, cited answers, cost/latency stats, and an 83-question golden set. The checked-in hybrid-rerank result reaches 0.952 document recall@10 and 0.711 exact-chunk recall@5.

Prosody Judge: LLM evaluation pipeline

Built an async eight-rubric LLM-as-a-judge tool with three-run self-consistency, structured outputs, uncertainty flags, checkpoint/resume behavior, and spend limits. The current four-item calibration is explicitly scoped as a smoke test pending human-correlation validation.

Experience

Independent AI/ML Engineer 2020-present

Build context-management, evaluation, fine-tuning, and local-inference pipelines across creative writing, code analysis, and image tooling. Prototype structured multi-model workflows across Claude, GPT, Gemini, and OpenRouter-backed models.

Technical Specialist, Interlock Industry August 2016-present

Install, calibrate, and troubleshoot embedded electronic devices against regulatory standards. Manage field-service work where reliability, documentation, and repeatable process are operational requirements.

Education and certifications

Self-directed applied ML study since 2016, including Stanford CS229, fast.ai Practical Deep Learning, and DeepLearning.AI coursework. Microsoft Azure AI Engineer Associate and Azure AI Fundamentals certifications.

Technical range

Python, TypeScript, React, FastAPI, MCP, LangChain, ChromaDB, PyTorch, Hugging Face, MLX, llama.cpp, Docker, GitHub Actions, Cloudflare Workers, evaluation design, retrieval systems, observability, LoRA/QLoRA.