Karina (Kai) Lark

AI / LLM Engineer | RAG, Evals

Saint Petersburg, Russia | open to relocation | Russian (native), English (C1)
icymirum@gmail.com | github.com/dontkaiad | t.me/hey_lark | heylark.dev

SUMMARY

AI/LLM engineer with almost 3 years of experience. I bring AI into production: support assistants, RAG systems, AI agents, knowledge-base search. I own quality control with evals and optimize costs through model routing and self-hosted inference.

I run projects end to end: architecture, development, deployment, monitoring, client communication. Tech lead and mentoring experience.

Before AI: 3+ years in technical support, monitoring and QA.

KEY RESULTS

- An LLM support assistant resolves ~73% of first-line tickets (600–800 per month) on its own across a 10,000+ device fleet, taking load off the support team.
- +43% bookings for a client on one project, driven by an LLM rollout.
- A personal knowledge-base assistant now finds every fact it needs: the share found rose from 72% to 100% on a test set after moving to hybrid search with query understanding.

SKILLS

AI Core: RAG (Qdrant, pgvector; hybrid FTS + vector search, RRF, LLM reranking), evals (blind-labelled golden sets, model A/B, recall@k, precision@k, regression evals in CI), grounding and LLM output validation, cost-aware model routing, prompt engineering, tool / function calling, structured output, self-hosted LLM (llama.cpp, Ollama), Claude / OpenAI / Gemini API, Whisper, vision prompting (multimodal LLMs)

Engineering: Python, FastAPI, PostgreSQL, Docker / Compose, Caddy, CI/CD (GitHub Actions), auth / SSO design, React / Vite, Linux / CLI, Git, REST API, Telegram Bot API

Delivery: scoping, technical documentation, backlog decomposition and ICE prioritization, roadmaps and sprints, production handoff

EXPERIENCE

Project Work — AI/LLM Engineer — Jun 2026 – Present

- Build and run production services (Docker Compose, Caddy, CI/CD): AI agents and Telegram bots with Mini Apps, RAG on pgvector, cost-aware Claude model routing, self-hosted LLMs (llama.cpp, Ollama).
- Lead projects from scratch to delivery: architecture, build, communication with team and clients. Monitoring with Langfuse, evals in CI.

Stack: Docker, Caddy, CI/CD, pgvector, Claude API, llama.cpp, Ollama, Langfuse, Telegram Bot API

Designflex — Technical Owner (AI/LLM Integration) — Feb 2026 – Jun 2026

Remote, product IT company

- Owned system architecture and technical direction, plus scope, delivery, quality and communication with clients and the team; led a team of 4 developers.
- Directed an embedding-model migration that raised knowledge-base answer quality from 0.59 to 0.82 on an internal score.
- Designed and implemented AI agents from scratch; technical owner of the LLM layer across several client projects.
- Built RAG systems on Qdrant (embeddings, semantic search); model selection and routing, LLM API integration (Claude, Gemini, OpenAI) and local models (Qwen), answer-quality monitoring via Langfuse.

Stack: Python, FastAPI, PostgreSQL, Qdrant, Langfuse, Docker, React

Project Work, NDA — AI/LLM Engineer Aug 2025 – Feb 2026

- Drove a 43% increase in bookings on one client project by introducing LLM features.
- Led projects from scratch to delivery: architecture, build, communication with team and clients.
- Designed and implemented LLM solutions for external clients: RAG systems on Qdrant, AI consultants and assistants with tool calling, Telegram bots with AI integration, cross-platform authentication, open-source data aggregation.

Stack: Python, FastAPI, Qdrant, Langfuse, PostgreSQL, Docker, React, GitLab CI/CD, llama.cpp, Ollama

Zvuk B2B — Engineer / AI Integration May 2024 – Aug 2025

Saint Petersburg (hybrid), B2B background-music platform

- Designed and shipped an LLM first-line support assistant for a fleet of 10,000+ devices: it autonomously resolves ~73% of L1 tickets (600–800 per month).
- Designed the prompt layer (system prompts, few-shot examples, response format) and structured the knowledge base as model input; tested answer quality and iterated on prompts.
- Introduced photo-based recognition of working vs. faulty Raspberry Pi units on the support line (multimodal LLM, vision prompting).
- Built fleet telemetry: requirements gathering, Grafana dashboards, Telegram alerting.
- Mentored a mid-level engineer across all work processes.

Stack: Python, Claude API, Linux, Grafana, Confluence, Jira

DatsTeam — Engineer / ML Integration Sep 2023 – Jan 2024

Moscow, B2C real-time platform

- Contributed to rolling out ML models on the first line of support: gathering and formalizing requirements, testing models, working with the development team.
- Took on-call incident duty: worked through incidents with developers, wrote postmortems, met SLAs.
- SQL analysis of user sessions, Grafana alerting, REST API diagnostics, product health and performance monitoring.

Stack: Linux, SQL, Grafana, Jira, REST API

QA and Technical Support 2019 – 2023

PROJECTS

Nexus & Arcana — Telegram AI agent system with a Mini App, in production. Two-stage intent classifier (11+ domains), cost-aware Haiku/Sonnet routing under regression tests, validate→retry→fallback. RAG memory: hybrid search (FTS + pgvector, RRF, Haiku rerank), answers checked for grounding. Model A/B comparison on cost and answer quality. Evals in CI.

Stack: Python, aiogram, FastAPI, PostgreSQL, pgvector, SQLAlchemy, Alembic, Claude API, Voyage AI, Whisper, React / Vite, Docker

github.com/dontkaiad/nexus-arcana

job-hunter-agent — Telegram bot + dashboard, in production. LLM job scoring: Haiku first pass, Sonnet judge for borderline cases. Calibration cut scoring bias from +13 to +1.6; quality is checked on a golden set of 115 blind-labelled vacancies. Pluggable sources, add-by-URL with SSRF guard, a live market-analysis dashboard.

Stack: Python, aiogram, Telethon, FastAPI, PostgreSQL, Claude API, APScheduler, React / Vite, Docker

github.com/dontkaiad/job-hunter-agent

Zarya (heylark Booking) — An LLM booking assistant: added to chats, holds a conversation, and helps navigate open slots and bookings. Web calendar in production: role-based access (guest / friend / admin) that hides booked slots from unauthenticated users, a free/busy engine, an ICS feed.

Stack: Python, aiogram, Claude API, FastAPI, PostgreSQL, SQLAlchemy, Docker

github.com/dontkaiad/nexus-arcana

EDUCATION

St. Petersburg State Institute of Film and Television, Advertising and Public Relations, 2023
Samara University, Foreign Philology, 2019