# Карина (Кай) Ларк

## AI / LLM-инженер | RAG, Evals

Санкт-Петербург, Россия | открыта к релокации | русский (родной), английский (C1)
icymirum@gmail.com | github.com/dontkaiad | t.me/hey_lark | heylark.dev

## ПРОФИЛЬ

AI/LLM-инженер, почти 3 года опыта. Внедряю ИИ в продакшен: ассистенты поддержки, RAG-системы, AI-агенты, поиск по базе знаний. Отвечаю за контроль качества с помощью evals, оптимизирую стоимость роутингом моделей и self-hosted инференсом.

Веду проекты целиком: архитектура, разработка, деплой, мониторинг, работа с клиентами. Опыт техлида и менторства.

До AI — 3+ года в технической поддержке, мониторинге и QA.

## КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ

- LLM-ассистент поддержки сам закрывает ~73% обращений первой линии (600–800 в месяц) на парке 10 000+ устройств и разгружает команду поддержки.
- +43% записей у клиента на одном из проектов благодаря внедрению LLM.
- Ассистент по личной базе знаний стал находить все нужные факты: доля найденных выросла с 72% до 100% на тестовом наборе после перехода на гибридный поиск и разбора вопроса перед поиском.

## НАВЫКИ

**AI Core:** RAG (Qdrant, pgvector; гибридный поиск FTS + вектор, RRF, LLM-реранк), эвалы (golden sets со слепой разметкой, A/B моделей, recall@k, precision@k, регрессионные эвалы в CI), grounding и валидация вывода LLM, экономный роутинг моделей, prompt engineering, tool / function calling, structured output, self-hosted LLM (llama.cpp, Ollama), Claude / OpenAI / Gemini API, Whisper, vision-промпты (мультимодальные LLM)

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

**Delivery:** скоупинг, техническая документация, декомпозиция и ICE-приоритизация, роадмапы и спринты, передача в продакшн

## ОПЫТ РАБОТЫ

**Проектная занятость** — AI/LLM Engineer — июнь 2026 – настоящее время

- Разрабатываю и сопровождаю production-сервисы (Docker Compose, Caddy, CI/CD): AI-агенты и Telegram-боты с Mini App, RAG на pgvector, cost-aware роутинг моделей Claude, self-hosted LLM (llama.cpp, Ollama).
- Веду проекты с нуля до деливери: архитектура, разработка, общение с командой и клиентами. Мониторинг через Langfuse, эвалы в CI.

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

**Designflex** — Technical owner (AI/LLM Integration) — февраль 2026 – июнь 2026

*Удалённо, продуктовая IT-компания*

- Отвечала за архитектуру и техническое направление проекта, а также за скоуп, деливери, качество и коммуникацию с клиентами и командой; вела команду из 4 разработчиков.
- Провела миграцию модели эмбеддингов: качество ответов по базе знаний по внутренней оценке выросло с 0.59 до 0.82.
- Спроектировала и внедрила AI-агентов с нуля; технический владелец LLM-слоя на нескольких клиентских проектах.
- Построила RAG-системы на Qdrant (эмбеддинги, семантический поиск); выбор и роутинг моделей, интеграция LLM API (Claude, Gemini, OpenAI) и локальных моделей (Qwen), мониторинг качества ответов через Langfuse.

**Стек:** Python, FastAPI, PostgreSQL, Qdrant, Langfuse, Docker, React

**Проектная занятость, NDA** — AI/LLM Engineer август 2025 – февраль 2026

- Обеспечила рост записей на 43% на одном из клиентских проектов благодаря внедрению LLM.
- Вела проекты с нуля до деливери: архитектура, разработка, общение с командой и клиентами.
- Спроектировала и внедрила LLM-решения для внешних клиентов: RAG-системы на Qdrant, AI-консультанты и ассистенты с tool calling, Telegram-боты с AI-интеграцией, кроссплатформенная авторизация, агрегация данных из открытых источников.

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

**Zvuk B2B** — Инженер / AI Integration май 2024 – август 2025

*Санкт-Петербург (гибрид), B2B-платформа фонового аудио*

- Спроектировала и внедрила LLM-ассистента первой линии поддержки для парка 10 000+ устройств: ассистент автономно решает ~73% обращений L1 (600–800 в месяц).
- Спроектировала промпт-слой (system prompts, few-shot примеры, формат ответов) и структурировала базу знаний как вход для модели; тестировала качество ответов и дорабатывала промпты.
- Внедрила распознавание исправных и неисправных Raspberry Pi по фото на линии поддержки (мультимодальная LLM, vision-промпты).
- Построила телеметрию парка устройств: сбор требований, дашборды Grafana, алертинг в Telegram.
- Менторила middle-инженера по всем рабочим процессам.

**Стек:** Python, Claude API, Linux, Grafana, Confluence, Jira

**DatsTeam** — Инженер / ML Integration сентябрь 2023 – январь 2024

*Москва, B2C real-time платформа*

- Участвовала во внедрении ML-моделей на первую линию поддержки: сбор и формализация требований, тестирование моделей, взаимодействие с командой разработки.
- Дежурила по инцидентам: разбирала их с разработчиками, писала постмортемы, соблюдала SLA.
- SQL-анализ пользовательских сессий, Grafana-алертинг, диагностика REST API, мониторинг работоспособности и производительности продукта.

**Стек:** Linux, SQL, Grafana, Jira, REST API

**QA и техническая поддержка** 2019 – 2023

## ПРОЕКТЫ

**Nexus & Arcana** — Telegram ИИ-агентная система с Mini App, в продакшене. Двухуровневый классификатор интентов (11+ доменов), экономный роутинг Haiku/Sonnet под регрессионными тестами, validate→retry→fallback. RAG-память: гибридный поиск (FTS + pgvector, RRF, реранк Haiku), ответы с проверкой grounding. A/B-сравнение моделей по стоимости и качеству ответов. Эвалы в CI.

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

github.com/dontkaiad/nexus-arcana

**job-hunter-agent** — Telegram-бот + дашборд, в продакшене. LLM-скоринг вакансий: Haiku первым проходом, Sonnet-судья для пограничных случаев. Калибровка снизила смещение скоринга с +13 до +1.6; качество проверяется на golden set из 115 вакансий со слепой разметкой. Подключаемые источники, add-by-URL с SSRF-guard, живой дашборд анализа рынка.

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

github.com/dontkaiad/job-hunter-agent

**Заря (heylark Booking)** — LLM-ассистент бронирования: добавляется в чаты, ведёт диалог и ориентирует по свободным слотам и записям. Веб-календарь в продакшене: ролевая модель доступа (guest / friend / admin) со скрытием занятых слотов для неавторизованных, free/busy-движок, ICS-фид.

**Стек:** Python, aiogram, Claude API, FastAPI, PostgreSQL, SQLAlchemy, Docker

github.com/dontkaiad/nexus-arcana

## ОБРАЗОВАНИЕ

СПбГИКиТ, Реклама и связи с общественностью, 2023

Самарский университет, Зарубежная филология, 2019