

Карина Ларк

ИИ / LLM-инженер

Санкт-Петербург, РФ (в разъездах) · удалённо · открыта к релокации · русский (родной), английский (C1)
icymirum@gmail.com · github.com/dontkaiaad · t.me/hey_lark

Профиль

ИИ / LLM-инженер: строю LLM-фичи и довожу до продакшена — RAG и семантический поиск, экономный роутинг моделей, контроль вывода (validate / retry / fallback) и self-hosted / локальный инференс. В продакшене мои собственные проекты — интент-классификация, экономный роутинг Haiku/Sonnet под регрессионными тестами, промпт-кэширование, постоянная память. За плечами 6 лет продакшн-бэкграунда (поддержка, QA, мониторинг). Ищу роль AI / LLM Engineer, где смогу владеть поведением и надёжностью LLM от и до (RAG, роутинг, контроль вывода и системы вокруг них) в развивающейся продуктовой команде.

Навыки

AI Core

Продакшн-RAG (pgvector / Qdrant, эмбединги) · экономный роутинг моделей · промпт-инжиниринг · контроль вывода (валидация, retry, fallback) · tool / function calling · структурированный JSON-вывод · self-hosted / локальные LLM (Llama, Qwen; llama.cpp, Ollama) · промпт-регрессия · маршрутизация интентов · оптимизация стоимости инференса · промпт-кэширование · LLM API (Claude, OpenAI, Gemini, OpenRouter) · Whisper · Vision

Engineering

Собственная production-инфраструктура: VPS, который я веду в одиночку от и до — Docker / Docker Compose, Caddy (reverse proxy / HTTPS), CI/CD (GitHub Actions), мониторинг контейнеров с heartbeat и cron-healthcheck'ами. Observability (Langfuse, Grafana, Graylog) · auth / SSO / SSRF-харднинг · PostgreSQL / SQL · React / Vite · FastAPI · REST API, вебхуки · Linux · CLI (bash, ssh) · Git · Python (чтение, ревью, правки — AI-assisted)

Delivery & Ownership

Скоупинг и сбор требований · техническая документация · быстрое прототипирование · передача в продакшн

В развитии

Эвал-пайплайны (offline golden sets, LLM-as-judge, regression gates) · беглость в async Python

Проекты

Nexus & Arcana — Telegram ИИ-агентная система (Python, в продакшене)

Двухуровневый классификатор интентов по 11+ доменам; экономный роутинг Haiku/Sonnet под регрессионными тестами; validate → retry → fallback с JSON-парсингом; постоянная память; Vision + Whisper. Arcana добавляет RAG поверх гримуара (база знаний), структур раскладов и контекста по каждому клиенту. Mini App (React + Vite + FastAPI).

github.com/dontkaiaad/nexus-arcana

job-hunter-agent — Telegram-бот + React-дашборд для автопоиска вакансий (Python, в продакшене)

LLM-скоринг с экономным роутингом против структурированного профиля кандидата; подключаемая Source-абстракция забирает Telegram-каналы и публичные job-API за единым интерфейсом; пайплайн add-by-URL прогоняет любую ссылку через полный скоринг с SSRF-guard. Self-hosted на одном VPS: PostgreSQL, Caddy / HTTPS, Telegram-login авторизация, healthcheck'и контейнеров и алертинг в Telegram.

github.com/dontkaiaad/job-hunter-agent

klgpff — Мультимодельный чат-роутер для креативного письма (Python)

Классификатор на Haiku маршрутизирует запросы между Opus / Sonnet / Haiku по типу задачи. Промпт-кэширование срезает стоимость ввода повторных ходов до ~10%; учёт стоимости по каждому чату с разбивкой input / output / cache; постоянная память фактов с семантическим поиском по pgvector.

github.com/dontkaiaad/klgpff-bot

Ещё проекты — на моём домене (hey_lark.dev): self-hosted Telegram-боты, Mini Apps и ИИ-агенты на инфраструктуре, которую веду сама.

Опыт работы

Интеграция ИИ в бизнес — Designflex

Февраль 2026 – июнь 2026 · удалённо · продуктовая IT-компания

Стек: Python, Gemini/OpenRouter, Claude, Qdrant, Langfuse, FastAPI, PostgreSQL, Telegram Mini Apps, React

Технический владелец LLM-слоя клиентских проектов: проектирование RAG-систем, тюнинг ретривала и эмбеддингов, промпты, контроль вывода, выбор и роутинг моделей, интеграция LLM API и мониторинг (Langfuse/Grafana).

- Спроектировала платформу поддержки и ИИ-аналитики с семантическим поиском по 4 потокам данных (Qdrant, 8 коллекций; Gemini через OpenRouter; мониторинг Langfuse); направила миграцию эмбеддингов — retrieval вырос до 0.82 (с 0.59).
- Спроектировала ИИ-коуч продаж и переговоров (голос → транскрибация → структурированный LLM-коучинг) и написала ТЗ; настроила промпты и роутинг моделей через OpenRouter.
- Спроектировала пересборку бота ассистент-платформы в единого ИИ-агента (структура команды, рейтинги, создание задач, контекстные ответы); настроила промпты и перенесла трекинг задач из Google Таблиц в веб.
- Делёвери на проектах: декомпозиция и ICE-бэклог, роадмап и спринты на разработчиков; ТЗ, сметы, приёмка.

Плюс ещё несколько клиентских ИИ/LLM-проектов в рамках роли — здесь самые показательные по стеку.

Фриланс и самообучение

Август 2025 – февраль 2026

- **Фриланс по интеграции ИИ в бизнес:** интегрировала LLM в существующие продукты и делала Telegram-ботов с нуля с ИИ-интеграцией для small-business и founder-клиентов (под NDA).
- **Интенсивное самообучение** по LLM, RAG, проектированию агентов и оптимизации стоимости инференса — фундамент для Nexus & Arcana; включая локальную LLM-песочницу на объединённых в сеть Raspberry Pi (локальная Llama).

PromptOps / QA — Zvuk B2B

Май 2024 – август 2025 · Санкт-Петербург (гибрид) · B2B-платформа фоновой музыки

Стек: Python, Claude/LLM API, SQL, Grafana, Usdesk, Confluence, Jira, SSH, RPi/Android

- Спроектировала промпт-часть и базу знаний LLM-ассистента поддержки, который автономно решает ~73% обращений L1 (600–800/мес); системные промпты, few-shot, база знаний как вход модели, тестирование ответов.
- Спроектировала телеметрию на парке 10к+ устройств; настроила дашборды Grafana и алертинг в Telegram.
- QA (smoke, регресс, A/B); прошивка edge-устройств по SSH; менторила middle-инженера.

Ранний опыт — поддержка, QA и мониторинг (2019–2024)

DatsTeam · Centrobite/agera.ru · Avito · Joom · стриминговая платформа (NDA)

- L2-поддержка и мониторинг на высоконагруженной real-time платформе (SLA отклика ~15 мин; Graylog, алерты Grafana → Telegram, диагностика REST API).
- L1–L3 поддержка + менеджмент проектов; CI со стороны поддержки (GitLab, Jenkins, Grafana); омниканальная B2B/B2C-поддержка; ручное QA и модерация; собственные расширения для Chrome и базы знаний.

Образование

Санкт-Петербургский государственный институт кино и телевидения — Реклама и связи с общественностью · 2022–2023

Самарский университет — Зарубежная филология (английский, немецкий язык и литература, с правом преподавания) · 2016–2019