Михаил Бальхаев

ML Engineer / Generative AI Engineer / Product Developer

Москва m.balkhaev@gmail.com +7 966 006-60-68 t.me/balkhaev balkhaev.com



ML-инженер в генеративном AI: диффузия для фото и видео, синтез речи, LLM-агенты и GPU-инфраструктура под них. Делаю продукты целиком, от пайплайна моделей и оценки качества до бэкенда и интерфейса: в 2026 году это сервисы генерации медиа, продукты на данных, CRM и игра на Unreal Engine. До 2024 года 13 лет в веб-разработке, в том числе в VK (Delivery Club) и Яндексе (Яндекс Еда).

Опыт

AIGF, luv.club, adorely.co, sinly.io — Март 2026 – сейчас

Generative AI Engineer

- Платформы AI-компаньонов (18+) с генерацией фото, видео и 3D, чатом и голосом: мультитенантный бэкенд на Hono, воркеры Redis/BullMQ, GPU-задачи на RunPod.
- Фото с сохранением внешности на Qwen Image 2.1: LoRA для согласованных правок, до четырёх референсов и автоматическая перерисовка, если лицо «уплыло».
- Видео: text-to-video и image-to-video с двухпроходным сэмплированием и интерполяцией RIFE до 48 fps; замена лица и всего персонажа на MiniMax H3 со swap-LoRA.
- 3D-сцена в видео: референсный рендер задаёт только первые 5 из 20 шагов сэмплирования, так что камера и композиция сохраняются, а движение остаётся естественным.
- Память компаньонов: извлечение фактов LLM с исправлениями, гибридный поиск по эмбеддингам OpenAI в pgvector, лексическое переранжирование и оценка релевантности LLM.
- Проверка возраста на сгенерированных медиа: детекция лиц, ViT-классификатор возраста и логистический пробник на эмбеддингах CLIP, обученный на 9 574 примерах. На отложенных данных верные отказы выросли с 283 до 350 из 400, ложные упали с 26 до 3 из 302.
- Агенты на Grok с инструментами: режиссёр студии настраивает съёмки и проверяет результат, 3D-редактор правит сцены с валидацией и контролем ревизий; REST и MCP API с OAuth 2.1.
- Голос: свой TTS на CPU (Kokoro-82M в ONNX, 54 голоса) и распознавание на Whisper.

Funcorp, Кипр — Ноябрь 2024 – март 2026

Senior Product Developer

- Начинал с поддержки основного продукта, потом делал новые AI-продукты с нуля.
- Диффузионные пайплайны для фото и видео (Wan 2.2, LTX, SD, SDXL) в ComfyUI и DiffSynth-Studio; сбор датасетов и обучение LoRA.
- Синтез речи: дизайн голосов и дообучение модели для естественного звучания.
- LLM-функции на Qwen3.5-9B, ChatGPT и Grok; RAG на Qdrant.
- Данные и бэкенд: Kafka, ClickHouse, PostgreSQL, Redis, Next.js, Hono, FFmpeg.

AI и медиа

Studio, film.balkhaev.com

Производство AI-фильмов: агент-продюсер на Claude превращает идею в сценарий и поставленные кадры, рендерит референсные проходы в Blender через MCP и генерирует кадры на H3. Дубль принимается, только если прошёл декодирование кадров, проверку визуальной моделью и транскрибацию речи.

MediaPipes, mediapipes.balkhaev.com

Генерация фото, видео и речи на ComfyUI и vLLM: Qwen Image 2.1 INT8 с десятью референсами, тайловый апскейл до 8 Мп, Higgs TTS 3 (4B). INT8-attention для H3 ускорил сэмплирование с 13,0 до 6,8 с/шаг (задача замены: со 129 до 80 с) и попал в прод после замеров сходства лиц, липсинка, LPIPS и слепого сравнения.

Persons

Сервис персонажей: семь референсных ракурсов и лист персонажа, проверка лиц на CPU (YuNet, 512-мерные эмбеддинги AuraFace), раскадровки до четырёх персонажей в кадре, очередь задач на PostgreSQL со SKIP LOCKED и справедливым распределением.

Modeler, modeler.balkhaev.com

Персонажи и превиз в Blender: image-to-3D (TRELLIS.2, Pixal3D) в GLB с текстурами, text-to-motion на риге из 63 костей с проверкой контакта стоп, реконструкция лица по снимкам Qwen.

Montage

Разбор видео: описания кадров от Qwen3-VL-4B, транскрипт WhisperX с выравниванием слов и склейки PySceneDetect в одном таймлайне, модели загружаются по очереди, чтобы уместиться в одну GPU; перекаст видео на H3.

Inference

Планировщик GPU для домашней RTX 5080 (ComfyUI, локальные LLM, Blender): допуск по измеренным пикам RAM и VRAM, совместный запуск только после трёх сопоставимых одиночных, прогретые модели и бюджеты, растущие после нехватки памяти.

Локальная LLM

Qwen3.8-Flash-Next (125B, GGUF) на одной RTX 5080 через Strata: эксперты распределены между VRAM, RAM и SSD, контекст 262k, около 100 токенов/с; шлюз с API OpenAI и Anthropic и маршрутизацией по аккаунтам.

Продукты на данных

ggpick.club

Поиск игр на русском, английском и китайском: данные Steam и сообществ в PostgreSQL, гибридный поиск на эмбеддингах multilingual-e5 в Qdrant с переранжированием cross-encoder и запасным путём при отказе моделей, транслитерация и исправление раскладки. 200+ уникальных посетителей в день.

Polywin

Исследование Polymarket: типизированные клиенты рыночных данных со сверкой стакана по WebSocket, сканер возможностей с учётом глубины и комиссий, разбор новостей LLM с проверкой цитат и прогнозы с оценкой по Brier score и калибровке. Сигнал требует 30+ закрытых рынков и статистически значимого преимущества.

AI News

Радар новых моделей и AI-инструментов из 44 источников (GitHub, Hugging Face, arXiv, Reddit, X): пересказы на русском с проверкой цитат, гибридный семантический поиск с cross-encoder на CPU, сигналы роста по звёздам и коммитам. Вынос блокирующей работы из event loop снизил p99-паузы с 585 до 62 мс.

FastNews

Новостная лента из RSS и X: безопасное извлечение статей, дедупликация по URL и тексту, обогащение LLM (сущности, темы, утверждения, прогнозы) с проверкой схемы и цитат, версионированный API для Polywin.

Soclogin

Сервис для собственных аккаунтов Instagram, X и Avito: хранилище учётных данных с AES-256-GCM, идемпотентные браузерные операции на Playwright, ленты для новостных продуктов и диагностика по MCP.

Сайты и заказы

CRM Земстрой

CRM продаж для компании по продаже земли: интеграции с amoCRM, Avito и ЦИАН на Redis/BullMQ, транскрибация звонков Whisper и их оценка LLM, бот первого ответа с версиями промптов, уведомления в Telegram о новых лидах и нарушениях SLA.

Алтай — живая земля, алтайземля.рф

Сайт посёлка с интерактивной 3D-долиной на Three.js: потоковая загрузка сцен на мобильных (первая загрузка с 7,2 до 3,7 МБ), лес из 79 125 деревьев по чанкам (с 5,9 до 1,2 млн треугольников), генплан на 568 участков общий для 2D и 3D, 31 индексируемая страница.

balkhaev.com

Падение в чёрную дыру Шварцшильда в браузере: трассировка лучей на WebGL в координатах Пенлеве–Гульстранда, линзированный каталог звёзд с доплеровским и гравитационным сдвигом, тесты против независимого численного интегрирования.

Инструменты и игры

DevHub

Платформа разработки для меня и моих AI-агентов: управление процессами с учётом зависимостей, процесс worktree и релизов с проверками, знания о проектах (Canvas и Kanban) через MCP, мониторинг деплоев Coolify.

Gamza

Изометрический экстракшен-автобаттлер на Unreal Engine 5.8 и C++: детерминированная headless-симуляция с проверяемыми реплеями, тактический ИИ отряда (слух, фланги, подавление), живой мир между рейдами, карты по геоданным и автоматическая кинематографичная камера.

Habar

Пиксельная idle-стратегия на TypeScript: симуляция без DOM, бой выстрел за выстрелом с фиксированным сидом, фракции, конвои и экономика, процедурный ландшафт, который рисуется в Web Workers.

Веб-разработка

Яндекс, Москва — Октябрь 2022 – август 2024

Разработчик интерфейсов

- Перешёл из VK вместе с командой Delivery Club в рекламное направление Яндекс Еды: аналитика и быстрая проверка рекламных гипотез.
- Основной webview Яндекс Еды и Яндекс Вендора, микрофронтендная архитектура.
- Рекламный раннер, первым партнёром которого стал Haval на старте в России; почти вся баннерная реклама, отладка аналитики и новые метрики.

VK, Delivery Club — Июнь 2020 – октябрь 2022

Ведущий разработчик

- Поддерживал delivery-club.ru, фактически как full-stack: админки для партнёров и юристов, рекламные форматы, внутренние UI- и API-библиотеки.
- После инцидента сделал для юристов систему документов с версиями и аудит-логами.

Data-screen, Москва — Июнь 2019 – март 2020

Fullstack-разработчик

- Сервис спортивных трансляций по высокоуровневым идеям, все технические решения на мне (Node.js, React, PostgreSQL, Redis, Docker); проект остановил ковид.

Смарт Солюшн, Москва — Ноябрь 2015 – декабрь 2018

Senior Frontend Developer

- Фронтенд десятка высоконагруженных сайтов из топ-50 Alexa: переписывание legacy, оптимизация и сборка; рекламные интеграции; обучение верстальщиков и код-ревью.

Oversun, Веб-разработчик — 2014 – 2015

Сайты на WordPress с нуля, свои модули, руководство младшим верстальщиком.

Ideegroupe, Сочи, Веб-разработчик — 2013

WordPress и ModX: модули для городского новостного портала, оптимизация SQL.

RedBox, Бурятия, HTML-верстальщик — 2011 – 2012

Шаблоны для Joomla и WordPress, jQuery.

Навыки

Модели	Qwen Image, H3, Wan 2.2, LTX, SD, SDXL, TRELLIS.2, Kokoro, Higgs TTS, Whisper, Qwen3-VL
ML-инженерия	Обучение LoRA, датасеты, оценка (сходство лиц, LPIPS, липсинк), CLIP-пробники, эмбеддинги и RAG, INT8 и GGUF
Сервинг	ComfyUI, vLLM, ONNX Runtime, RunPod, планирование GPU, Redis/BullMQ
LLM и агенты	Tool use, MCP, Claude, GPT, Grok, Qwen, pgvector, Qdrant
Бэкенд и данные	Python, TypeScript, Node.js, Bun, Hono, FastAPI, PostgreSQL, ClickHouse, Kafka, Redis, Qdrant, SQLite, Docker
Фронтенд и 3D	React, Next.js, Three.js, WebGL, Unreal Engine 5 (C++), Blender, Playwright

Детали

Языки	Русский (родной), английский (выше среднего, B2)
Образование	Среднее
Формат	Полная или частичная занятость, проекты; удалённо, гибрид или офис; готов к переезду по России.