

Matt Kerkstra

AUSTIN, TX 918-398-3588 MATTKERKSTRA@GMAIL.COM GITHUB.COM/MKERKSTRA LINKEDIN.COM/IN/MATT-KERKSTRA

SUMMARY

Staff-level platform engineer. Seven years building production ML infrastructure and the systems other engineers run on. #1 of 45+ contributors across five repositories at my current employer. 2,076 commits, 1.5M LOC over three years. Built the Kubernetes ML platform, the model-serving stack, a voice-to-clinical-note pipeline, and a graph-based clinical AI application that shipped from architecture to production in sixteen days. Architectural bets compound: a single 2023 Postgres/PostGIS migration is still enabling new product surface in 2026.

CORE SKILLS

ML / LLM	Kubeflow, KServe, vLLM, LangGraph, Langfuse, RAG, Milvus, BAAI/bge-m3, Presidio, AWS Bedrock, Anthropic Claude, MCP
INFRASTRUCTURE	Kubernetes, Istio (ambient mesh), ArgoCD, Helm, Kustomize, Terraform, Karpenter, GitHub Actions, cert-manager, external-secrets
DATA	PostgreSQL, PostGIS, Aurora, ClickHouse, Snowflake, Drizzle ORM, Kysely, SQLAlchemy, Mongo -> PG migrations, Redis / Valkey, BullMQ, Kafka
LANGUAGES	TypeScript, Python, Go, SQL
APP STACK	NestJS, Fastify, FastAPI, Pydantic, React, Next.js, Jotai, React Query, OPFS, Service Workers
OBSERVABILITY	Grafana, Tempo, Alloy, OpenTelemetry, Mimir, Loki, Pyroscope

EXPERIENCE

Senior Software Engineer at VideHealth · Remote

MAY 2023 - PRESENT

#1 contributor of 45+ engineers across 5 repositories. Operating at Staff scope.

Architected and shipped a production AI clinical-note templating system from empty repo to production in **16 days** across 3 services. Graph-based LLM pipeline with Postgres-backed checkpointing, semantic section matching via BAAI/bge-m3 embeddings + Milvus, dual-detector PHI anonymization (Presidio + LLM), SSE streaming, and an admin review UI with structured rich-text editing.

Built the production **Kubernetes ML platform** from an empty repo. Istio ambient mesh, ArgoCD app-of-apps with Kustomize overlays, KServe + vLLM for self-hosted model serving (embeddings, ASR), Langfuse for LLM observability, Milvus for vector search. 13 namespaces / 9 services. Cost stayed flat; new services deploy with a Helm chart and an ArgoCD app. Six engineers now contribute regularly.

Led **MongoDB -> PostgreSQL/PostGIS** migration of a clinical analyses data model (100M+ records); cut heavy queries from >5s to <500ms. Unlocked a multi-year cascade of product capabilities - spatial segmentation storage, per-patient overlays, and the clinical recommendation engine in production today. Own every layer of that chain, from schema through Python algorithms to TypeScript integration across 10+ versions.

Built an end-to-end **voice-to-clinical-note pipeline**: offline-resilient browser capture (OPFS, Service Workers), Jotai recording state, BullMQ async processing, transcription, structured LLM summarization, multi-language support. Migrated the service into the ML cluster for direct in-mesh access to inference infra. 60K+ TS LOC across client and API.

Delivered the unified-appointments backend that landed the **largest dental services organization (DSO) contract** in the United States - vault practice search, dual-mode practice support, optimizations for legacy PMS integrations.

Established a **Kubeflow**-based experimentation framework standardizing data versioning and model promotion across the ML org.

Run the bi-weekly **Backend Guild** (18+ months) driving cross-team architectural alignment. Drove org-wide adoption of typed query patterns (Kysely + footgun-prevention bots) and AI-assisted developer tooling - first mover on Cursor rules, CLAUDE.md, and MCP integrations a year before mainstream.

Software Engineer at Paperspace · Remote

JUN 2022 - MAR 2023

Rebuilt Node.js + Stripe billing to support **2.5x YoY revenue growth** while keeping payments under 200 ms.

Implemented real-time fraud and sanctions checks that shut down illicit GPU crypto-mining from embargoed regions, **reducing chargebacks >50%**.

Senior Software Engineer / Software Engineer at Hotel Engine · Remote

JUL 2021 - JUN 2022

Introduced bundle splitting, CDN routing, and feature flags, shrinking **mean deploy time from 15 min to 6 min** for 40+ engineers.

Drove Redux -> React Query migration, reducing cold-start data fetches 40% and bundle size 20%.

Software Developer at Reynolds & Reynolds · Houston, TX

FEB 2019 - JUN 2021

Converted a 20-year-old version-control system for F&I forms from VB6/SQL to COBOL + Pick BASIC - responsible for distributed delivery, usage tracking & billing, and integration with F&I systems. Assumed lead role two weeks after onboarding and delivered on schedule.

SIDE PROJECTS

Narrative Nexus at **narrative.sh** · AI-powered companion for tabletop RPG dungeon masters. Co-built with a group of long-time friends who are also software devs.

TYPESCRIPT · NEXT.JS · GO

EDUCATION

