Krishan Saini

Principal AI Engineer · LLM & Retrieval Systems · Cloud Platform

reachout@krishansaini.com · +91 98888 89957 · krishansaini.com · linkedin.com/in/krishansaini · toptal.com/developers/resume/krishan-saini · Remote · India (UTC+5:30) · works US & EU hours

SUMMARY

Engineer with 12+ years in production systems, the last three spent building AI-native products end to end. Shipped a legal-research platform with retrieval over 19.4M court judgments and a clinical voice product built to HIPAA Security Rule controls that transcribes, summarises and medically codes physician calls. Comfortable owning the whole path: retrieval and prompt design, the Python and Node services behind them, the React/TypeScript surface on top, and the Kubernetes underneath. Prior platform work at Booking Holdings and Pixis means the systems survive real traffic.

TECHNICAL SKILLS

AI / LLM: LLM integration (OpenAI, Anthropic) · RAG & hybrid retrieval · vector search (pgvector, OpenSearch, Pinecone) · embeddings & chunking strategy · prompt engineering & evals · agent / tool-use design · citation grounding · ASR (Whisper) · LangChain · prompt caching & latency tuning

Languages: Python · TypeScript / JavaScript · SQL · Bash

Backend: FastAPI · Node.js / Express · REST & GraphQL · Postgres · Redis · OpenSearch / Elasticsearch · MongoDB · Pub/Sub · Pulsar · async workers · microservices

Frontend: React · Next.js · TypeScript · Tailwind · design systems · browser extensions

Cloud: GCP (GKE, Cloud Run, BigQuery, Pub/Sub, Vertex AI, Cloud Functions) · AWS (EKS, ECS, Lambda, S3, DynamoDB, API Gateway)

Platform / DevOps: Kubernetes · Docker · Terraform · Ansible · GitHub Actions · Cloud Build · Jenkins · GitOps

Reliability & Security: Prometheus · Grafana · ELK · Stackdriver · chaos testing · HA design · HIPAA · GDPR

EXPERIENCE

Principal Engineer · AI Products · Thinkscoop Technologies — 2023–Present

Independent consultancy; also contracting through the Toptal network

CourtMesh.ai: AI legal research for Indian litigators

- Built and operate an AI legal-research platform over 19.4M Supreme Court and High Court judgments plus 49,487 Acts and 213,540 statutory sections: ingestion, embeddings, retrieval, API, and the React/TypeScript product on top.
- Designed the retrieval pipeline end to end: document parsing, chunking strategy, embedding generation, and hybrid search combining vector similarity with metadata filters over OpenSearch, tuned by reranking to lift top-k relevance over a keyword baseline.
- Built the case-research chat layer with citation grounding against the judgment corpus so the model cannot fabricate case law, plus response caching that cut LLM spend and tail latency.
- Instrumented the funnel end to end: GA4 and conversion tracking, UTM attribution, and a marketing analytics console that attributes signups back to individual campaigns.
- Own production reliability: Kubernetes deployment, autoscaling for embedding and inference workers, and observability on query latency and retrieval quality.

OnCall.ai: clinical voice product built to HIPAA Security Rule controls

- Shipped a product built to HIPAA Security Rule controls that transcribes, summarises and medically codes physician–patient phone calls, working the full path from audio ingestion to the physician-facing review UI.
- Built the speech-to-text, summarisation and medical-coding pipeline on top of LLMs, with prompt design and evaluation tuned for clinical accuracy and reviewer trust rather than raw benchmark scores.
- Hardened the data path for HIPAA: encryption in transit and at rest, audit logging, access controls, and explicit PHI-handling boundaries between services.
- Built the React/TypeScript clinician dashboard for review, edit and sign-off, a human-in-the-loop workflow where the AI drafts and the physician stays accountable.

Other builds

- Built an autonomous outbound GTM system: lead sourcing and enrichment, LLM scoring with a score-before-reveal guard, sequenced email, and automatic reply handling behind safety gates that hold sensitive threads for a human.
- Built a job-aggregation engine that normalises listings from heterogeneous XML feeds with an LLM, then serves semantic search, saved searches and alerts over MongoDB.
- On client engagements, write the design docs the client's team builds from (model selection, latency-versus-accuracy tradeoffs, retention and rollback policy) and review PRs in their repos instead of adding process.

Principal DevOps Engineer · Pixis — 2021–2023

AI-driven advertising and marketing platform

- Ran the cloud platform behind Pixis' AI products in production: model-serving infrastructure on GKE, Pub/Sub-driven async pipelines, and BigQuery analytics supporting fast iteration on ML features.
- Broke a legacy monolith into microservices on GCP with Pub/Sub and Kubernetes, improving fault isolation and unblocking parallel feature work across product teams.
- Automated CI/CD on Cloud Build and GitHub Actions so engineers shipped to production daily, with canaries and auto-rollback as guardrails.
- Managed petabyte-scale BigQuery pipelines powering campaign analytics and downstream ML training; centralised logging and metrics on ELK and Prometheus.

Lead DevOps & Cloud Architect · Booking Holdings — 2018–2021

Travel and hospitality

- Designed and scaled GKE clusters serving millions of travel searches a day, including the search and personalisation services that fronted ranking models.
- Led the migration of core infrastructure from AWS to GCP, cutting operational cost by ~22% while holding uptime through the cutover.
- Replaced manual provisioning with Terraform and Cloud Build, putting infrastructure under version control and making changes repeatable across environments.
- Stood up observability on Prometheus and Stackdriver and cut median incident response time by ~40% through better alerting and runbook coverage.

Senior Cloud Consultant (DevOps & SRE) · VMware — 2016–2018

Hybrid cloud and SaaS migrations

- Migrated SaaS workloads from AWS to VMware's GCP hybrid cloud with zero downtime, working alongside customer engineering teams through each cutover.
- Wrote the internal DevOps playbooks adopted by 100+ engineers across global teams and drove adoption of TDD and CI pipelines.
- Built automated disaster-recovery procedures and tested them under load so they actually held when called on.

Software Engineer · web & backend · Earlier roles — 2014–2016

- REST API and backend development across web and mobile platforms, real-time tracking, routing optimisation and payment-gateway integrations.

EDUCATION

Bachelor of Engineering, Computer Science · Panjab University, Chandigarh — 2014