Zardar Khan

zardar.khan@icloud.com | linkedin.com/in/zardar-khan | github.com/khankanz | blog
Toronto, ON

Summary

Applied Research Engineer specializing in clinical NLP systems. 4+ years shipping production LLM pipelines in healthcare: structured report generation (constrained decoding), clinical RAG, PHI de-identification. Track record of rapid execution – concept to production in weeks, paper to implementation in days.

Core Skills

LLMs & NLP: Hugging Face Transformers • llama.cpp • RAG pipelines • Constrained decoding (Outlines/XGrammar/GBNF) • Knowledge distillation • Model quantization
Infrastructure: FastAPI • Docker • AWS/Azure • Postgres • FHIR/OMOP
Core: PyTorch • Python • Pydantic • SQL

Selected Projects

Constrained Decoding (CRANE-style) for Structured Medical JSON 2025 — Present

- Built constrained decoding pipeline converting free-text pathology reports to schema-valid CAPeCC (DCIS Resection) JSON; achieved 100% schema validity on TCGA and internal pilot samples
- Implemented free-reason → switch-token → JSON window generation; ported from Hugging Face to llama.cpp (GGUF) for large-model inference on A100s
- Diagnosed and patched Outlines → llama.cpp logits incompatibility breaking FSM-based token masking; enabled cross-stack constrained generation
- Approved to scale to 17K reports; optimizing inference via XGrammar port (targeting sub-1 min from current 1-3 min/report)

Clinical RAG — Caring Contacts 2024

- Collaborated with Psychiatry to build RAG system generating personalized post-discharge hope letters from patient charts
- Implemented proposition-level chunking + two-stage retrieval (BM25 → Qwen reranker) with clinician-defined safety guardrails
- Shipped concept to production in 3 weeks; system now in clinical trial (ISBD 2025 poster presentation)

Professional Experience

Sunnybrook Research Institute Sept. 2023 – Present
Research Software Engineer *Toronto, ON*

- Built FHIR data extraction pipeline scaling from 20K to 2M+ clinical reports; reducing batch processing time from 120 days to < 4 hours (700x speedup) by implementing temporary indexing tables in Postgres.
- Led PHI de-identification from 45% to 75% F1 using synthetic data generation + GPT-4 knowledge distillation; mentored student who extended approach to 90% F1 on held-out test set.
- Shipped clinical RAG system in 3 weeks from concept to production; generates post-discharge communications for psychiatric patients using proposition-level chunking + two-stage retrieval (BM25→Qwen reranker); deployed with clinician-in-the-loop safety guardrails, now in clinical trial.
- Quantized RadBERT microcalcification classifier from 450ms → <100ms (4.5x speedup) preserving 95% accuracy; work accepted for Oral Presentation at IGTxIMNO 2024.

gojitech Aug. 2021 – Sept. 2023
Software Engineer *Toronto, ON*

- Reduced prototype feedback cycle from 2+ weeks to 1.5 weeks through on-site clinical shadowing and workflow redesign.
- Diagnosed & resolved Chrome 2GB tab memory limitation affecting voice recording; implemented streaming audio compression, eliminating data loss for long clinical sessions.
- Rebuilt clinical transcription from batch (30s max, 5-10s latency) to real-time streaming; implemented WebSocket architecture with Azure STT delivering live intermediate → final transcript refinement as users speak

Techna Institute (Health Informatics) Nov. 2018 – Nov. 2020
Business Application Analyst *Toronto, ON*

Education

Toronto Metropolitan University Toronto, ON
Bachelor of Engineering, Biomedical Engineering

Publications & Presentations

Khan Z et al. *AI-Personalized Caring Contacts for Psychiatric Discharge.* IASR International Summit, Boston — 2025 — Poster

Khan Z et al. *AI-Enabled Suicide Prevention via Caring Contacts.* ISBD Annual Conference, Chiba, Japan — 2025 — Poster

Khan Z et al. *RadBERT Microcalcification Classifier (95% accuracy).* IGTxIMNO 2024 — Oral Presentation