Zardar Khan

Toronto, ON | zardar.khan@icloud.com | github | linkedin | portfolio

Summary

I build ML systems in healthcare—messy data, tight timelines, real consequences. 4+ years shipping production pipelines: constrained decoding (180x speedup), clinical RAG (3 weeks to trial), PHI de-identification (45% → 90% F1). I reverse-engineer undocumented systems and make them work.

Work

Constrained Decoding for Medical JSON | *code* 2025

- Pathology reports → structured JSON. Prompt-only approaches hit ∼67% validity—1 in 3 broken at scale.
- Built two-stage pipeline: reason first, then constrained emit. Monkey-patched xgrammar into llama.cpp (27x), then vLLM migration with continuous batching (7x more).
- 180x speedup (3 min → 1 sec). 100% schema validity. 17K reports in 5 hrs. Scaling to 1M+.

Caring Contacts RAG 2024

- Personalized letters of hope for patients following psychiatric discharge that reference specific details from their stay.
- BM25 → reranker. Proposition-level chunking beat doc/paragraph for retrieval. Prompt guardrails for safety.
- Concept to clinical trial in 3 weeks. Presented ISBD 2025 (Japan), IASR 2025 (Boston).

NLP to OMOP Pipeline | *slides* 2024

- Radiology reports → NLP classification → FHIR R4 → OMOP tables. End-to-end.
- Extended fhir.resources + fhirpy with custom Breast Radiology IG profiles. Terminology layer maps NLP outputs to RadLex/SNOMED codes. Async Aidbox client with upsert semantics.
- Batched inference pipeline (DeBERTa classifiers) with HuggingFace datasets. OMOP server populated from live radiology reports.

Mammography Classification | *code* 2024

- Sort DICOMs by artifact type (mag view, biopsy tool) when metadata was too inconsistent to rely on.
- EfficientNetB0 + active learning loop: 50 labels → 90+% accuracy on 400 samples.

PHI De-identification 2024

- Off-the-shelf NER models miss hospital-specific patterns—too generic for internal data.
- Built augmented data pipeline: SQL to extract real PHI patterns → GPT-4 generates plausible text with XML placeholders → swap with real identifiers → train NER.
- 45% → 75% F1. Supervised student who extended to 90+% with decoder model.

FHIR Extraction 2023

- Needed 2M+ clinical reports. Existing queries took 120 days.
- Reverse-engineered undocumented Observation → DiagnosticReport structure. Temporary Postgres indexes.
- 120 days → 4 hours (700x). Found 500K missing reports from import errors.

Jobs

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

- ML engineer embedded across research groups. Built all pipelines above. Go-to for pathology/radiology report extraction—wrote custom parsing to stitch fragmented reports that appear line-by-line.
- Routinely pull and create datasets with thousands of reports for downstream research. Work adopted by 2 external labs.
- Quantized RadBERT classifier: 450ms → <100ms (4.5x), 95% accuracy. Oral presentation at IGTxIMNO 2024.

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

- Real-time transcription (WebSocket + Azure STT, replaced 30s batch). Medication extraction (GPT-Neo).
- Fixed Chrome 2GB memory bug killing long recordings.

Side Projects

Bookkeeping Agent — Built tool-use agent from scratch to analyze my bank statements. Hit context limits on large files, calculator failed on OCR errors, pivoted to DuckDB for SQL on CSVs. Found I was spending $70/mo on Starbucks.

Pronunciation App — Speech-to-IPA transcription, LLM-as-judge grading against target pronunciation, daily iterations. Customer base of 1, retention 100%.

Education

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

Publications

Khan Z et al. *AI-Personalized Caring Contacts.* IASR 2025 (Boston), ISBD 2025 (Japan).

Khan Z et al. *RadBERT Microcalcification Classifier.* IGTxIMNO 2024 — Oral Presentation.

Skills

LLMs & NLP: Transformers · llama.cpp · vLLM · RAG · Tool-use Agents · Constrained Decoding · NER

ML Techniques: Quantization · Active Learning · Knowledge Distillation · Synthetic Data Generation

Stack: PyTorch · FHIR · OMOP · Postgres · Docker · AWS/Azure · FastAPI