# Khushi Keswani

+91 9879791980 | khushikeswani97@gmail.com | linkedin.com/in/khushi-keswani-b3471b28b | github.com/KhushiKeswani

## EDUCATION

**Shree Swaminarayan Institute Of Technology** — Ahmedabad, Gujarat
*Bachelors in Computer Engineering* — *Aug. 2023 – May 2027*

## EXPERIENCE

**AI and ML Lead** — *August 2025 – May 2026*
*Google Developer Groups*

- Led Generative AI and ML initiatives for Google Developer Groups (GDG), designing and delivering hands-on workshops that upskilled 100+ students in practical AI tools and industry-relevant skills.
- Directed Google Cloud Study Jams sessions, running structured cloud infrastructure training and mentorship for the student community.

**AI and Cloud Intern** — *May 2025 – June 2025*
*AICTE in collaboration with Microsoft*

- Built a customer-churn prediction pipeline for a telecom use case, preprocessing 7,043 customer records and benchmarking Decision Tree vs. Random Forest models through 5-fold cross-validation.
- Applied SMOTE to mitigate class imbalance and evaluated generalization using precision/recall, selecting Random Forest with 84% cross-validation accuracy.

## PROJECTS

**Mamba-Based Sequence Model — HDFS Log Anomaly Detection | Python, PyTorch, NumPy, CUDA** — *Sept 2026*
github.com/KhushiKeswani/mamba-training-from-scratch

- Implemented Mamba's selective state-space mechanism from scratch — Δ/B/C projections, discretization, causal conv, gated residuals, selective scan — validating NumPy ↔ PyTorch equivalence to $<6\times10^{-6}$ error.
- Reformulated the recurrent scan as an associative prefix-scan and compiled it for GPU execution on an NVIDIA T4, verifying gradient flow across all 4 blocks end-to-end.
- Trained on highly imbalanced HDFS logs (~3% anomalies), reaching 95.0% test F1, 98.3% ROC-AUC, and 96.2% PR-AUC on a 10K-sample held-out set.

**RepoGraph | Python, AST, NetworkX, sentence-transformers, PyVis** — *July 2026 – Present*
github.com/KhushiKeswani/repograph

- Built a GraphRAG system turning raw Python repos into queryable knowledge graphs: AST-based entity extraction, NetworkX call/dependency graph, and hybrid retrieval combining semantic embeddings with graph traversal.
- Designed scope-aware call resolution disambiguating same-named functions across files — distinguishing local, cross-file, and unresolved/external calls instead of naive string matching.

**Conversational Chatbot | Python, FastAPI, Gemini API, JWT, Redis, Docker** — *March 2026 – May 2026*
github.com/KhushiKeswani/chatbot

- Built a conversational AI backend with FastAPI, JWT authentication, and async SQLAlchemy; fixed three blocking synchronous calls (bcrypt, Gemini API, Redis) freezing the event loop under concurrent load via asyncio.to_thread.
- Load-tested with Locust to 42.5 req/s at 99.86% reliability, 110ms median latency, and a 97.7% Redis cache hit ratio.
- Containerized with a multi-stage Dockerfile and built a GitHub Actions CI/CD pipeline, debugging 13+ pipeline failures (env-variable mismatches, missing Redis service container, cache-key normalization) to reach a stable automated test suite.

**Vox — A CPU-Optimized Streaming Voice Assistant | Python, Whisper, Silero VAD, Gemini API, Piper TTS** — *Sept 2026*
github.com/KhushiKeswani/vox

- Built a real-time CPU-only voice assistant (Silero VAD → Whisper ASR → streaming Gemini → Piper TTS), using a producer-consumer queue to decouple audio capture from inference on dual-core, no-GPU hardware.
- Cut speech-boundary detection error from 276ms to 19ms across 10 annotated recordings; profiled 100 end-to-end turns (P50 12.5s, P95 16.0s, P99 17.8s) and traced tail latency to Whisper ASR (P99 9.4s vs. 1.6s median) and Piper TTS (P99 10.9s vs. 3.3s median) spikes, while Gemini stayed stable (P99 3.8s).

## TECHNICAL SKILLS

**Languages:** Python, SQL (Postgres)

**AI/ML:** PyTorch, Hugging Face Transformers, Mamba/State-Space Models, RAG & GraphRAG, sentence-transformers, NetworkX, Whisper

**Backend & APIs:** FastAPI, Gemini API, JWT Authentication, Async SQLAlchemy, RESTful APIs

**Infra, Data & Testing:** Docker, GitHub Actions (CI/CD), AWS (EC2), Redis, CUDA, Pandas, NumPy, scikit-learn, Locust