Aahil Khan

B.E. Computer Engineering
Thapar Institute of Engineering & Technology

aahilminookhan@gmail.com
linkedin.com/in/aahil-khan
github.com/aahil-khan
aahil-khan.xyz

Skills

Languages: Python, TypeScript, JavaScript
Frameworks: React, Next.js, Node.js, Express.js, FastAPI, Flask
Databases: PostgreSQL, MySQL, Redis
Cloud & DevOps: AWS (EC2, Lambda, API Gateway, DynamoDB, S3), Docker, Git, GitHub Actions
AI/ML: LLM APIs, RAG pipelines, vector embeddings, semantic search
Testing: Jest, Vitest

Professional Experience

Research Intern — Samsung PRISM *Remote*
PyTorch, SpeechLLMs (SALM, Qwen-Audio) | Accent-Invariant Representation Learning Jul 2026 – Present

- Researching **parameter-efficient accent adaptation** for multilingual SpeechLLMs — LoRA adapters, MoE dynamic adapter routing, and accent-aware front-end encoders — to improve ASR/AST robustness on accented speech without full model retraining.
- Benchmarking baseline SpeechLLM pipelines (SALM, Qwen-Audio, SALMONN) and running **accent failure analysis** to identify weak accents and adaptation targets.
- Work targets **publication of research findings** in a peer-reviewed conference/journal as the closure deliverable.

Full Stack AI Intern — Recruit by Zariya AI *Bangalore, India (Remote)*
Next.js, TypeScript, Node.js, AWS S3, LiveKit | Interviews by Zariya AI Apr 2026 – Present

- Delivered features across frontend, backend, and real-time services for a live AI interview platform, working closely with recruiter and candidate feedback.
- Built **audio based interview evaluation pipelines** using AWS S3 and Gemini Files API, producing structured scores and recruiter facing report data.
- Built recruiter facing pipeline tooling with server-side filtering for candidate workflow and status management.

Team Lead & Backend Engineer — Thapar University *Patiala, India*
Node.js, PostgreSQL, Redis, Docker, AWS EC2, Flask, MySQL Oct 2024 – Jul 2025

- Led a team of 3 engineers to build **Thapar EduTube**, a lecture archival platform serving **15,000+ students**; shipped a CLI sync tool with automated YouTube uploads, **reducing upload times significantly.**
- Evaluated Elasticsearch vs PostgreSQL FTS and migrated search infrastructure, reducing operational complexity while maintaining performance under **k6 load testing.**
- Separately re-architected the university **Datesheet Generator** from session based state to a persistent, multi user MySQL backed workflow, **reducing scheduling errors by 40%.**
- Integrated with Oracle PeopleSoft to automate structured data ingestion and export between legacy and modern systems.

Selected Projects

SOP Opera — Agentic Industrial Safety Intelligence Platform
Python, TypeScript, FastAPI, LangGraph, PostgreSQL/pgvector, Next.js, Docker

- Built a compound-risk detection engine combining gas sensor readings, work permits, isolation state, and worker location to catch hazards conventional single sensor thresholds miss, cutting the false negative rate from **44.5% to 0%** across 593 labeled cases and alarming **28 minutes** earlier.
- Designed a **LangGraph multi agent pipeline** with conditional agent fan out and hybrid pgvector/SQL retrieval over a clause level regulatory corpus, so every generated citation is validated against retrieved evidence and the risk verdict stays fully deterministic.
- Built a durable Postgres job queue and a non-blocking WebSocket layer with per client backpressure, with a **hash chained tamper evident audit trail** for every recorded safety decision.

FlowSync — AI Project Memory for Coding Agents
TypeScript, Python, AWS (Bedrock, Lambda, DynamoDB, S3), MCP, Next.js

- Designed an AI project memory system using **event-driven architecture** to extract decisions, risks, and tasks from git diffs via Bedrock models and persist them for coding agents via MCP.
- Developed a VS Code extension and MCP server for context logging, project search, and branch aware RAG retrieval.
- Implemented **branch aware RAG** with DynamoDB caching, achieving **1.26s median push-to-search latency** (p95: 1.60s).

Konta — Context-Aware Browsing System *(Samsung PRISM Web Agent Hackathon Winner)*
TypeScript, React, IndexedDB, Transformers.js, ONNX

- Built a local first Chrome extension that converts browsing activity into an on-device knowledge base using **on-device semantic embeddings** for private, low latency search.
- Designed a **knowledge graph linking system** connecting pages using semantic, temporal, and contextual relationships.
- Engineered sessionization logic with context boundary detection to prevent unrelated browsing tasks from merging.

Achievements

Winner — Samsung PRISM Web Agent Hackathon *2025*
Innovation Award — Agentic AI Hackathon, Ulster University (UK) *2025*

Education

Thapar Institute of Engineering & Technology *CGPA: 9.16*
Bachelor of Engineering, Computer Engineering Aug 2023 – Present

Indian Institute of Technology Madras *CGPA: 8.34*
Diploma in Programming (concurrent online CS diploma, run alongside the B.E.) Sept 2023 – Sept 2025