

SHRESHTTH KUMAR AGARWAAL

AI/ML Engineer | M.Sc. Student, AI for Industrial Applications | Amberg, Germany
agarwalshreshth3@gmail.com | [linkedin.com/in/shreshtth-kumar-agarwaal](https://www.linkedin.com/in/shreshtth-kumar-agarwaal) | github.com/Blackbird-3

SUMMARY

AI/ML engineer focused on taking generative AI from prototype to production: real-time voice agents, retrieval-augmented generation, knowledge graphs and agentic workflows, backed by solid full-stack and data engineering foundations. Experience spans AI engineering at Sftwtrs.ai, GenAI audit automation at GlobalLogic, data analytics at KPMG India and enterprise software at MetLife. At Sftwtrs.ai I also worked on the business side, acquiring and onboarding a large infrastructure client through pre-sales, stakeholder discovery and POC delivery, which fuels my interest in entrepreneurship and in building AI products that solve real customer problems. Currently pursuing an M.Sc. in Artificial Intelligence for Industrial Applications at OTH Amberg-Weiden.

EXPERIENCE

AI Research Engineer

Jun 2025 – Jan 2026

Sftwtrs.ai

Gurugram, India

- Acquired and onboarded a large infrastructure company as a client: led pre-sales, stakeholder discovery with the CIO and CHRO, and end-to-end POC delivery.
- Built a real-time AI hiring automation POC with an interactive AI avatar for candidate screening, using WebRTC, Next.js and Python.
- Developed an enterprise-grade AI voice automation platform integrating Tata SmartFlo telephony with LiveKit and SIP, reducing average call handling time by 35% through a low-latency WebSocket architecture.
- Designed a compliant healthcare conversational AI with Dialogflow CX on Google Cloud Platform (GCP), implementing secure session management and authentication for patient interactions.
- Managed production infrastructure with Docker, tmux and self-hosted services; designed the PostgreSQL (Supabase) backend with serverless RPC functions for secure schema management.
- Conducted R&D on low-code knowledge graph architectures, delivering a POC with LangGraph and FalkorDB that validated visual domain modeling for non-technical users.
- Worked in rapid agile sprint cycles to ship production-ready code, run user-acceptance testing (UAT) and maintain CI/CD pipelines.

Generative AI Intern

Jan 2025 – May 2025

GlobalLogic

Noida, India

- Developed the MVP of AuditAI, a generative AI audit automation platform that improved audit efficiency by 50%–70% through scalable GenAI workflows, real-time follow-ups and editable question banks.
- Performed extensive prompt engineering for LLM-based scoring, maturity assessment and summary generation, reducing bias and increasing audit consistency.
- Built a full-stack React + FastAPI application with spaCy-based anonymization for PII protection, on a domain-configurable, scalable architecture with human-in-the-loop review for quality and compliance.

Data Analysis, Visualization Intern

Jun 2024 – Aug 2024

KPMG India

Gurugram, India

- Designed and executed ETL pipelines to extract, transform and clean multi-source enterprise data.
- Built 5+ Power BI dashboards that made complex customer and sales datasets accessible to stakeholders, using DAX queries, measures and calculated columns.
- Developed proofs of concept leading to an estimated 20% improvement in decision-making efficiency.

Full Stack Development Intern

May 2023 – Jul 2023

MetLife Global Shared Services

Noida, India

- Developed a budgeting tool that automatically accrues monthly team budgets based on headcount, matches expenses from PeopleSoft and provides real-time budget visibility.
- Wrote numerous Microsoft SQL Server (T-SQL) stored procedures powering budget allocation, expense matching and reporting.
- Built the Angular front-end and the .NET MVC (C#) API layer on top of the SQL Server back-end.
- Gathered functional requirements with stakeholders in an Agile/Scrum setup and contributed to sprint planning, accelerating development cycles by 30%.

PROJECTS

Autonomous Voice & Call Automation Platform

Jul 2025 – Sep 2025

Gemini Realtime API, LiveKit, WebRTC, SIP, FastAPI, Supabase

- Designed a production-ready, low-latency voice pipeline connecting commercial telephony (Tata SmartFlo) via a SIP trunk into automated LiveKit agent rooms.
- Integrated Gemini's native realtime speech-to-speech models over persistent WebSockets, bypassing separate STT/TTS stages for fluid, sub-second conversational latency.
- Configured turn detection (barge-in handling) and noise cancellation, reducing average call handling time by 35%; tracked multi-turn conversation state with serverless PostgreSQL RPC functions.

AuditAI: Enterprise Audit Automation Platform

Jan 2025 – May 2025

React, FastAPI, LangChain, Groq, Presidio, spaCy, Docker

- Built an optimized RAG pipeline to automate evidence verification for regulatory compliance, accelerating audit lifecycles by 50%–70%.
- Orchestrated multi-model inference with LangChain on low-latency Groq endpoints and evaluated cost savings with self-hosted Llama 3 models.
- Engineered an inline PII anonymization module (Microsoft Presidio, spaCy) in FastAPI applied before any third-party LLM call, plus a React human-in-the-loop review interface for compliance officers.

Autonomous Home Assistant Robot (NAO Simulation)

Apr 2026 – Jun 2026

Python, NAOqi SDK, Choregraphe, WebSockets

- Engineered a schedule synchronization and pre-reminder pipeline for a simulated NAO humanoid robot, combining asynchronous task scheduling with real-time motion control.
- Mapped calendar events to gestures, localized audio alerts and collision-free navigation paths via Choregraphe and the NAOqi Python API, triggered by stateful WebSocket event listeners.

Sound Slice: Music Source Separation App

Oct 2023 – Dec 2023

Flutter, Dart, Firebase, Replicate API, Python

- Built an end-to-end deep learning inference pipeline in a cross-platform Flutter app that isolates vocals, drums, bass and instrumentals from arbitrary audio files.
- Secured user-uploaded audio with Firebase Storage and token-based client authentication.

EDUCATION

OTH Amberg-Weiden

Amberg, Germany

M.Sc. Artificial Intelligence for Industrial Applications

Mar 2026 – Expected Mar 2028

BML Munjal University

Gurugram, India

B.Tech. Computer Science and Engineering, Specialization in AI/ML

Aug 2021 – Jul 2025

- Relevant courses: Machine Learning, Artificial Intelligence, Deep Learning, Generative AI & LLMs, Explainable AI, Data Structures & Algorithms

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, C/C++, Java, C#, SQL (T-SQL), Dart, HTML/CSS

AI Engineering: Agentic Workflows, Multi-Agent Systems, RAG, GraphRAG, Vector Embeddings, Prompt Engineering, LiveKit, WebRTC, Dialogflow CX, PII Anonymization (Presidio, spaCy), Cypher Queries

Frameworks & Libraries: PyTorch, TensorFlow, LangChain, LangGraph, FastAPI, Flask, Next.js, React, Node.js, Express, Angular, .NET MVC, Flutter

Databases & Knowledge Graphs: PostgreSQL, Supabase, MS SQL Server, Neo4j, FalkorDB, Pinecone, Firebase Firestore

Cloud & DevOps: AWS (EC2, Lambda, ECS, API Gateway, CloudWatch), Google Cloud Platform, Docker, Terraform, Git, GitHub Actions, CI/CD, Linux

Data & BI: ETL Pipelines, Power BI (DAX, Measures, Calculated Columns)

Ways of Working: Agile/Scrum, Pre-Sales & Stakeholder Discovery, Cross-Functional Collaboration

AI Developer Tools: Claude Code, Cursor, GitHub Copilot, v0, Perplexity Pro

LANGUAGES

English: C1 (TOEFL) · German: B1 (currently learning) · Hindi: Native