Bengaluru, Karnataka, IN
ameybhavsar10@gmail.com

Amey Bhavsar

Software Development Engineer II

github.com/ameybh
linkedin.com/in/amey-bhavsar

Objective: Product Engineer with 3 years building AI/ML systems end-to-end. Skilled at shipping features fast, from customer discovery to production deployment. Seeking Product Engineer or Full-Stack roles in Germany.

Technical Experience

Tessact **Bengaluru, India**
Software Development Engineer II *January 2026 — Present*

- Built and shipped an AI-powered video repurposing platform end-to-end as sole engineer in **2 months**. Pipeline takes 2-hour podcasts and produces **30-40 branded clips** (30-60 seconds each) ready for social media. Replaces **4 weeks of manual editing with 1 hour**, delivering **95% ready-to-post quality**. Processed **500+ hours** for enterprise POCs.
- Designed **microservices architecture** with **FastAPI** service (**TessactAI**) using **PydanticAI** for **LLM orchestration** and **Django** backend. Built **DAG-style pipeline** orchestrating 11 jobs via **Celery** and **RabbitMQ**: transcription, face tracking, speaker resolution, clip generation, enhancement generation, and rendering. Each service independently deployable and scalable.
- Built complete **full-stack product**: **React** frontend with video upload/preview, **Django** backend with job orchestration, **Remotion** rendering on **AWS Lambda** (30 seconds per clip, parallel processing). Containerized face tracking service with **Docker** on **GCP Cloud Run** for auto-scaling. Set up **CI/CD** with **GitHub Actions**, **Claude Code** PR reviews, and **Ruff** pre-commit hooks.
- Optimized costs and quality: A/B tested **OpenAI GPT-5** vs **Google Gemini 2.5 Flash**, switched to Gemini for better quality at **60% lower cost**. Switched transcription from AWS ($1.44/hr) to **ElevenLabs Scribe v2** (**$0.40/hr**, **72% reduction**) with improved accuracy. Podcast output generation costs **$5 per hour** vs **$1,000 market rate** for manual editing.
- Built complex face tracking system combining **InsightFace** with frame-to-frame tracking and ReID for multi-speaker podcasts. Implemented talk score detection, head tilt tracking, focus state detection, and occlusion handling. Optimized to run only on final clip time ranges, saving processing time. Developed brand kit automation using LLMs to generate custom **Remotion** components matching client branding.
- Secured enterprise POCs with European and US brands including **Garena Free Fire** (livestream analysis POC). Created documentation (**Slack Canvas**, **API docs**), conducted knowledge transfer sessions, and pair-programmed with team members. Managed delivery using **Linear** with weekly sprint cycles in 6-engineer team at $2M funded startup.

Software Development Engineer *July 2022 — December 2025*

- Built face detection and tracking service replacing **AWS Rekognition**. Reduced cost from **$6 to $0.15 per hour** and improved processing speed. Saves **$2,500 per month**. Runs in production for clients including **JioHotstar** and **SunTV**. This work led to promotion in January 2026.
- Designed video editing automation system using **Whisper** (speech recognition), **TransNet-V2** (scene detection), **CLIP** (image analysis), and **Tesseract** (text detection). **LLM agents** analyze multimodal signals to plan cuts automatically. Reduced editing time from **60 minutes to 10 minutes** per video.
- Built video transcoding infrastructure using **FFmpeg**, **RabbitMQ**, and **Celery** with **NVIDIA GPUs**. Creates multi-resolution videos (1080p, 720p) with **HLS** and **DASH** streaming. Implemented **AWS MediaConvert** failover for reliability.
- Rebuilt product frontend using **Next.js 14** and **TypeScript**. Reduced crashes and improved code quality. Set up monitoring with **Sentry** and established coding standards for the team.

Skills

AI/ML & LLMs	OpenAI GPT-4/5, Google Gemini, PydanticAI, LangChain, PyTorch, TensorFlow, InsightFace, Whisper, CLIP.
Backend & APIs	Python, Django, FastAPI, Node.js, Celery, RabbitMQ, PostgreSQL, Redis, RESTful APIs, Microservices.
Frontend	TypeScript, React.js, Next.js 14, Remotion, Tailwind CSS.
Cloud & DevOps	Docker, Kubernetes, GitHub Actions, CI/CD, GCP (Cloud Run, Artifact Registry), AWS (Lambda, MediaConvert).
Tools & Practices	Git, Linear, Agile/Scrum, Pair Programming, Code Review, API Documentation, Ruff, Pre-commit Hooks.
Communication	English (**IELTS Band 8, CEFR level C1**), Hindi, Marathi, German (A1 - Learning).

Education

Bachelor of Engineering in Computer Engineering 2019 - 23
University of Mumbai, India CGPA: **9.15** / **10** (**1.4** German GPA)
Relevant Coursework: Deep Learning, Machine Learning, Neural Networks, Natural Language Processing Data Mining
Thesis: "Saathi – An AI Companion": Mental Health Chatbot for Emotion and Mental Disorder Detection using Deep Learning

Projects

Saathi - AI-Powered Mental Health Companion **July 2022 — June 2023**
Bachelor's Thesis Project | *Published IEEE Conference (INSCIRD-2023)*

- Developed end-to-end AI system for mental health support using deep learning on large-scale real-world data, achieving **4th Place at IIT Bombay Research Conclave (RESCON)** national competition.
- Web scraped and preprocessed **100,000+ Reddit posts** from mental health forums (Depression, Anxiety, Bipolar Disorder) for training robust classification models.
- Designed and implemented **multi-layer CNN architecture** for mental illness detection: **127-dimensional word embeddings** → CNN layer (ReLU activation) → 64-node CNN layer → MaxPooling1D → Flatten → Dense(250, ReLU) → Dense(3) output layer, achieving accuracies of **86.76%** (Depression), **89.79%** (Anxiety), and **95.51%** (Bipolar Disorder).
- Implemented **Seq2Seq encoder-decoder architecture with LSTM** for natural language generation, trained on CounselChat dataset (therapist-client Q&A pairs) to generate contextually relevant, empathetic therapeutic responses.
- Integrated **RASA framework** for Natural Language Understanding (NLU) with intent classification, entity extraction, and dialogue management to create production-ready conversational AI system.
- Engineered **multi-task learning system** performing simultaneous disease detection (3 classes) and emotion detection (8 classes) from user messages for comprehensive mental health assessment.
- Deployed Flask-based REST API for real-time model inference with React and Material UI frontend, enabling accessible mental health support via web and mobile platforms.
- **Published paper**: Saathi - An AI Companion INSCIRD-2023 (IEEE Students' Branch, Shri Sant Gajanan Maharaj College of Engineering, Shegaon), June 2023.
- **Technologies:** Python, TensorFlow, Keras, CNN, LSTM, Seq2Seq, Word Embeddings, RASA Framework, NLP, Flask, React, Material UI, Multi-task Learning, Deep Learning.

Achievements & Activities

Competitive Programming: Codechef rating **1821** (Ranked **104 globally**); Codeforces rating **1444** (Ranked **1315 globally**); ICPC Gwalior-Pune Regionals Rank **314 & 1139** (2020, 2021); Meta Hacker Cup Round 2 (2022).

Research & Hackathons: 4th Rank at IIT Bombay Research Conclave (RESCON) National Level (2023); Winner - CodebuG at ACM VIT, CodeMania, Knapsack 2.0, Algorithm Design, Web Design at SLRTCE (2020-2022).

Standardized Tests: GRE - Verbal: **169/170**, Quantitative: **165/170**, AWA: **4.5/6**; IELTS: **Band 8** (CEFR C1).

Leadership: Web Development Lead - Student Council; Technical Lead - Google Developer Student Club SLRTCE (2022-2023).