SHAHIDULLA VANTELA

AI/ML Engineer | Python | LLMs

Yemmiganur, Andhra Pradesh, India
prince.shahidulla007@gmail.com | +919666685485
LinkedIn | GitHub

PROFESSIONAL SUMMARY

AI/ML Engineer with 2+ years architecting production AI systems in document processing, NLP, and generative AI. Expert in scalable ML pipelines using GCP, AWS, LLMs (Gemini 2.5 Pro, BERT), and distributed frameworks. Track record: 70% cost reduction, 90% search improvement, 95%+ accuracy.

PROFESSIONAL EXPERIENCE

DataToBiz (AI Venture) – Mohali, India | AI Developer & ML Engineer | July 2025 - Present

• TeacherTech (AI Exam Grader):

- Built microservices platform (Flask + Celery + Redis + AWS S3) processing 100+ concurrent exams with 95% accuracy
- Multi-stage pipeline: Document AI OCR → Gemini 2.5 Pro structuring → PyMuPDF splitting → LLM grading
- Overcame API limits via chunking (>15 pages, >40MB), ensured reliability through JSON schemas
- 85% time reduction (48hrs vs 2wks), enabling scalable semester-end processing

• Voice AI Bot & AWS Infrastructure:

- Built conversational system (Whisper-1 → Custom LLM → gTTS) achieving <500ms latency with Retell AI
- Designed AWS S3 lifecycle policies reducing storage costs by 60% with 99.9% availability

Chervic Artificial Sciences – Bangalore, India | Artificial Intelligence Intern | March 2024 - Feb 2025

• Invoice Processing: Built extraction pipeline (PyMuPDF + Tabula) processing 500+ invoices monthly with 98% accuracy, reducing manual entry by 70% ($50K+ savings)

• Semantic Search: Deployed vector search (Elasticsearch + SBERT) on 100K+ products achieving 90% relevance improvement, 12% conversion uplift, <200ms latency

• BERT QA Chatbot: Fine-tuned BERT on 5K+ QA pairs (95% accuracy) with Whisper voice, automating 80% of queries via Gradio (6hrs daily savings)

Exposys Data Labs – Bangalore, India | Data Science Intern | June 2023 - Aug 2023

• Built XGBoost models achieving 95% accuracy on 500K+ records; created Plotly/Seaborn dashboards for stakeholder insights

Technocolabs Softwares – Indore, India | AI Developer Intern | July 2022 - Oct 2022

• Developed LSTM on 3M Spotify records achieving 95% skip prediction; collaborated on traffic mortality reduction using ensemble ML

KEY PROJECTS

TeacherTech - AI Exam Grading System | July - Nov 2025

Python | Document AI | Gemini 2.5 Pro | Flask | Celery | Redis | AWS S3

• **Architecture:** Flask API + Celery workers + Redis queue + AWS S3 storage with multi-stage pipeline (OCR → JSON structuring → student splitting → LLM grading)
• **Challenges:** Solved concurrency (100+ parallel jobs), context management (>15 pages, >40MB chunking), reliability (JSON schemas preventing hallucinations)
• **Impact:** 95% accuracy, 85% time reduction (48hrs vs 2wks), enabling scalable semester-end assessments

Voice AI Chatbot | Sep - Nov 2025

Whisper-1 | gTTS | Retell AI | Custom LLM | Flask

• **Pipeline:** Whisper-1 STT → Custom LLM reasoning → gTTS TTS achieving <500ms latency for real-time interactions
• **Evolution:** Prototyped with Flask, scaled to production using Retell AI framework
• **Impact:** Enabled hands-free voice interactions on company website, improving accessibility and engagement

Semantic Search Engine | Oct - Dec 2024

Elasticsearch | SBERT | Streamlit | KNN | Kibana

• **Implementation:** Vector search with SBERT embeddings on 100K+ products, Elasticsearch with KNN (<200ms latency), Streamlit frontend with Kibana monitoring
• **Impact:** 90% relevance improvement vs keyword matching, 12% conversion uplift, 10K+ concurrent users

BERT QA Chatbot | June - Oct 2024

BERT | Whisper | Gradio | Python

• **Development:** Fine-tuned BERT on 5K+ QA pairs (95% accuracy) with Whisper speech recognition for voice-based querying
• **Impact:** Deployed via Gradio, automated 80% of support queries, reduced response time by 6hrs daily

---

TECHNICAL SKILLS

Programming: Python (Expert), SQL (Advanced), R, Java, C++

AI/ML: TensorFlow, PyTorch, Scikit-Learn, XGBoost, LightGBM, Keras, BERT, Transformers, NLP

LLMs & Generative AI: Gemini 2.5 Pro, BERT, SBERT, Whisper-1, gTTS, Prompt Engineering, RAG, LangChain

Backend & APIs: Flask, FastAPI, Django, Celery, Redis, REST APIs, Streamlit, Gradio, Retell AI

Cloud & Infrastructure: Google Cloud (Document AI, Vertex AI), AWS (S3, EC2), Docker, Linux, Git

Data Engineering: PyMuPDF, Tabula, Pandas, NumPy, PostgreSQL, MongoDB, Elasticsearch, Neo4j, KNN Search

Visualization: Plotly, Matplotlib, Seaborn, Tableau, Kibana

Tools: GitHub, VS Code, Jupyter Notebook, Postman, Ngrok

EDUCATION

Master of Science in Artificial Intelligence | Heriot Watt University, Edinburgh, UK | Jan 2022 - Sep 2023
Coursework: Data Mining, Neural Networks, Deep Learning, Human-Robot Interaction, Big Data Management

Bachelor of Technology in Computer Science Engineering | Lovely Professional University, Punjab, India | Jul 2016 - Sep 2020
Specialization: Algorithms, Data Structures, Database Management, Python Programming

---

CERTIFICATIONS & ACHIEVEMENTS

• Google Cloud Platform (Document AI, Vertex AI) | HackerRank 5-star Python | LeetCode 200+ problems
• Production ML systems (99.9% uptime) | Active AI/ML community contributor