Varun Bikkumalla

Hyderabad, India | varunbikkumalla968@gmail.com | linkedin.com/in/varunc9 | github.com/varunb-ai

Summary

AI Engineer with 2+ years of experience building LLM-powered applications, AI agents, and data pipelines using Python. Experienced in voice AI, document classification, and LLM inference using vLLM, with a strong focus on evaluation, data quality, leak-safe benchmarking, and production reliability.

Experience

AI Developer, Forage AI, Hyderabad *Mar 2026–Present*

Private-equity document classification (8 document types plus "Other") for a global asset manager
Python, PyTorch, FinBERT, Qwen2.5-32B, vLLM, FAISS, UMAP, Azure ML (4×A100 80GB)

- Raised unseen-document accuracy from **49.51%** for an earlier FinBERT model (below a 61% keyword-rules baseline) to **89.08%** by fixing class imbalance and noisy labels and adding a binary "Other" gate.
- Traced that model's 99.62% validation score to **73% duplicated training data**; rebuilt a 57,361-document leak-safe corpus, cutting the validation–unseen gap from 50 points to **6** (97.06% vs 91.15%).
- Moved to **Qwen2.5-32B** on **vLLM**, which reads the non-English documents FinBERT misclassified: **94.75%** on a QA benchmark of 2,113 documents at 0.8 s/doc batched on 4×A100. Wrote the production infra requirements.
- Audited benchmark labels with the QA lead, fixing 55 and 37 wrong gold labels (the first benchmark re-checked in full, not only disagreements); with the models unchanged, scores moved from 91.15% to **92.82%** and 94.75% to **96.50%**.

Outbound voice AI agent: provider-directory verification POC for a US healthcare-data client
Python (asyncio, WebSockets), OpenAI Realtime API, Twilio Voice, Google Sheets API, pytest, pyright, CI

- Built and ran an agent that navigates phone menus and holds to verify listings with front desks: 1,388 calls to 903 providers in nine nights; **359 calls succeeded** against a target of 200, validating **358 providers (39.6%)**.
- Raised the per-call success rate from **18.5% to 32.6%** while growing volume from 47 to 260 calls a night, by replaying each night's failed calls offline on a simulated clock and fixing the cause in code before the next night.
- Enforced behavior in code rather than the prompt: audible narration fell from **42.9% to 6.9%** of tool turns, and saving answers from the live transcript cut yes/no replies from 2.8 s to **0.92–1.27 s**.
- Ran the POC for about $260: **$0.73 per validated provider**, $0.10 per connected minute. Guarded it with a 7,509-check offline suite (mutation testing caught 52 of 52 mutants) and consent-by-state, do-not-call and office-hours gates.

Associate Data Engineer, AionOS, Hyderabad *Jul 2024–Feb 2026*

- Owned **TrustMate**, the data-quality module of IntelliStream (internal data platform, team of 3): LLM-based anomaly detection, PII monitoring and remediation alerts using Gemini and DeepSeek.
- Contributed pre-load and post-load schema validation to scheduled Snowflake–Supabase ingestion, and natural-language dataset exploration to InsightsMate (Gemini, LLaMA; answers in 10–15 s end to end).
- Built the platform's AWS deployment pipelines with Docker and Kubernetes.

Frontend Developer Intern, S7Works, remote *Jul–Dec 2023*

- Built React.js features and a HubSpot analytics integration for two production products, Bitwork.ai and Workahub.

Projects

Stratax AI: AI interview practice platform (personal project) | strataxai.in | github.com/varunb-ai/Stx-BE
FastAPI, PostgreSQL, Redis, Qdrant, LangChain, Groq, Gemini, faster-whisper, Judge0, Docker

- Built evidence-linked scoring: four weighted criteria, each feedback item anchored to a quote or paraphrase of the answer; criteria with no input (e.g. delivery on a coding answer) are dropped and weights renormalized, not scored as zero.
- Built resume-driven mock interviews with local speech-to-text (faster-whisper), LangChain hybrid retrieval (Qdrant plus BM25) for company-specific questions and a Judge0 code sandbox; 234-test pytest suite.

Technical Skills

Languages: Python, SQL, JavaScript (React.js) | **LLMs:** Qwen models, OpenAI, Gemini
Serving & inference: vLLM, multi-GPU tensor parallelism, batch inference
Voice: OpenAI Realtime API, Twilio Voice, IVR handling
Evaluation: Benchmark design, leakage detection, label auditing, mutation testing, pytest
ML & retrieval: PyTorch, Hugging Face Transformers, FinBERT, LangChain, FAISS, Qdrant, BM25, UMAP
Backend & tools: FastAPI, PostgreSQL, Redis, Snowflake, Azure ML, AWS, Docker, Kubernetes, Claude Code

Education, Certifications & Achievements

B.Tech., CMR Institute of Technology, Hyderabad — GPA 8.96 / 10 *2020–2024*

- Databricks: Data Engineer Professional, Data Analyst Associate (2024); EDW and CDW Migration Delivery Labs (2025).
- CodeChef 3-star (highest rating 1624). Led a six-member team at BPR&D's national Manthan event.