

# Jayesh Bankoti

[jayeshbankoti@gmail.com](mailto:jayeshbankoti@gmail.com) · +91 84483 59710 · [LinkedIn](#) · [Website](#) · [GitHub](#)

## PROFESSIONAL SUMMARY

---

AI Engineering Leader with 8 years of experience (4+ at Staff level) building and scaling production AI systems across NLP and GenAI. Led organizations of up to 9 engineers, owning architecture, hiring, and roadmap for narrative intelligence platforms processing millions of social posts. Proven record of translating research-grade models into production systems that delivered 120% engagement growth and 82% infrastructure cost reduction, working directly with executive leadership on AI strategy.

## LEADERSHIP & STRATEGY

---

- Served as de facto Head of AI post-Series A at Logically AI, owning AI strategy, architecture, hiring, and execution across DS/ML tracks while reporting directly to the CEO.
- Built the Media Intelligence (NLP) division from zero, scaling from 0 to 3 engineers, then unified multiple divisions into a single 9-member AI/DS organization.
- Overhauled the NLP technical hiring process end-to-end and led all staff-level hiring across multiple company growth phases.

## EXPERIENCE

---

### **Tiger Analytics (Remote)** — *AI Solutions Specialist*

*Mar 2026 – Present*

- Built a cloud migration assessment framework using Treesitter AST extraction and Neo4j knowledge graphs that deterministically generates service profiles, IaC mappings, database inventories, and code-change reports across 9 migration patterns spanning AWS, GCP, Azure, and on-prem.
- Built a hybrid chatbot with an LLM router that dynamically invokes Table-Augmented Generation (DuckDB SQL over structured assessment reports), vector RAG, or both depending on query type, enabling structured and semantic queries over migration findings.
- Built a code migration agent for AWS EMR-to-GCP modernization handling diverse source artifacts (EMR Python scripts, Teradata/BTEQ, PySpark, shell scripts, cluster configs), using a generator-reviewer pattern with an independent LLM-as-judge validating generated code against source logic, dual-gated with dry-run execution until both pass.
- Built a post-migration rule validator using Dataplex-generated rules reviewed by an independent LLM judge to verify correctness before sign-off.

### **Logically AI** (Bengaluru, Remote)

*Dec 2021 – Mar 2026*

#### *Senior Staff Engineer – AI/DS Organization Lead*

*Sep 2023 – Mar 2026*

- Led the 9-engineer AI/DS organization as Senior Staff, the highest-scope technical role post-Series A, owning roadmap, architecture, quarterly delivery, and direct C-suite alignment.
- Drove 120% platform engagement growth by replacing E5 with Qwen3 LLM-based embeddings, GPU-accelerated UMAP and HDBSCAN clustering (Optuna-tuned), and MinHash-LSH deduplication to surface coordinated inauthentic posting patterns across millions of social posts.
- Reduced embedding-generation latency by 35% by productionizing inference with the PyTorch 2.x compiler stack, Sentence-Transformers, FlashAttention, and BF16 precision execution.
- Reduced infrastructure costs by 82% by consolidating microservices into a monolithic workflow and shifting orchestration from reserved to on-demand compute, backed by PostgreSQL state and messaging.
- Productionized the AI platform through a strategic migration to Databricks with robust CI/CD (Asset Bundles and service principals), aligning DS, engineering, and MLOps to accelerate narrative product delivery.
- Designed a modular agentic chatbot using SmolAgents (ReAct) with LangChain tools (search, SQL) and a Mosaic AI vector DB for RAG-backed narrative Q&A, with multi-turn analyst workflows via summarization-based memory.
- Engineered hybrid data infrastructure combining PostgreSQL (OLTP), Delta Lake (OLAP), and Elasticsearch to enable sub-second narrative retrieval and large-scale analytics.

#### *Staff Applied NLP Engineer – AI Division Head*

*Dec 2021 – Sep 2023*

- Built the Media Intelligence NLP team from 0 to 3 engineers, owning hiring, mentoring, technical direction, and full delivery of the AI/NLP roadmap.
- Built the Narrative Intelligence System that anchored the company's \$24M Series A raise (Vitruvian Partners, Amazon Alexa Fund), combining narrative detection, large-scale annotation pipelines, and analytics.
- Reduced model inference latency by 45% via ONNX Runtime and TensorRT with BF16 optimization, establishing it as the production deployment standard across the AI team.
- Built and scaled annotation pipelines (NER, sentiment, language detection, translation, geolocation) processing all social posts end-to-end, owning team delivery and system performance.

### **Brane Enterprises (Hyderabad)** — *NLP Team Lead*

*Nov 2020 – Dec 2021*

- Supervised a 12-member NLP team, owning design, integration, and delivery of NLP/ML modules spanning information extraction, classification, and question answering.
- Built an FAQ QA system (BM25 with RoBERTa fine-tuned on SQuAD 2.0) reaching 80% accuracy in internal evaluation, integrated into the production help-desk chatbot.

**VMock (Gurgaon)** — *Data Scientist – NLP & Deep Learning*

*Jul 2018 – Oct 2020*

- Built and deployed a BiLSTM-CNN resume classifier (F1 90%) across 4 live products used by 150+ universities, including Stanford, NTU, and London Business School.
- Constructed a 2,000+ node skill knowledge graph with graph-algorithm ranking for personalized job recommendations at 90% relevancy, powering the Wall of Jobs product across 30+ colleges.
- Developed job classification and recommendation models (CNN with collaborative filtering), deployed across the core university-facing jobs platform.

## SELECTED GENAI SYSTEMS (OPEN SOURCE & APPLIED RESEARCH)

---

### Autonomous AI Research System (Multi-Agent Pipeline)

[\[Video\]](#) — [\[GitHub\]](#)

- Built a LangGraph multi-agent pipeline with a deterministic Python supervisor controlling all routing (no LLM-based orchestration), ensuring predictable execution and bounded compute.
- Engineered an adversarial two-agent debate loop (Research and Critic) that iteratively critiques, defends, and re-scores research papers until convergence or an iteration limit.
- Decomposed the post-approval pipeline into isolated single-task SLMs (synthesis, hypothesis, writeup, evaluation) with clean context windows and fault-isolated retries.

### Hierarchical RAG Architecture for Financial Analysis

[\[Video\]](#) — [\[GitHub\]](#)

- Architected a hierarchical semantic RAG pipeline that clusters 150–200 token fragments into ~1,500 token thematic superchunks via HDBSCAN, capturing global context across non-contiguous pages of large financial corpora.
- Engineered FAISS (HNSW) retrieval with LLM-based query reformulation for acronym disambiguation, achieving stable, citation-grounded generation with sub-second latency.

### Multi-Persona Aligned LLM – Potterhead GPT

[\[Video\]](#) — [\[GitHub\]](#) — [\[HuggingFace\]](#)

- Fine-tuned GEMMA with QLoRA via two-stage training: domain grounding on all 7 books, then instruction tuning (SFT) on 50K+ multi-persona Q&A pairs, yielding accurate persona-switching and context retention.
- Deployed a streaming inference pipeline enabling token-wise generation with dynamic persona switching across character styles.

## PUBLICATIONS

---

Jayesh Bankoti, et al. (2022). *Logically at the Constraint 2022: Multimodal Role Labelling*. Co-authored paper, ranked **2nd** among 9 final systems from 100+ registered teams. [\[Leaderboard\]](#)

## TECHNICAL SKILLS

---

- **GenAI Systems:** LLM fine-tuning (QLoRA, PEFT), instruction tuning, RAG (graph, hierarchical), agentic AI (LangGraph, LangChain), TRL (SFTTrainer), context engineering, prompt harness engineering (CLAUDE.md, skill files, hooks, subagents), SmolAgents, FAISS, Neo4j
- **ML/DL & NLP:** PyTorch, Transformers, scikit-learn, SentenceTransformers, cuML, Optuna, HDBSCAN, UMAP
- **MLOps & Inference:** ONNX Runtime, TensorRT, FastAPI, Docker, CI/CD (Databricks Asset Bundles, service principals), MLflow
- **Data & Infrastructure:** Mosaic AI, PostgreSQL, Elasticsearch, MongoDB, Redis, Delta Lake, PySpark, Databricks
- **Other Tools:** Treesitter, Dataplex, Git, Flask, Doccano, Streamlit, TensorFlow

## EDUCATION

---

### Indian Institute of Technology, Delhi

B.Tech in Mechanical Engineering  
KVPY Fellowship Awardee

*Jun 2014 – May 2018*

GPA: 7.54/10  
July 2013