Harshit B.

Boston, MA | 857-565-4643 | brahmbhatt.h@northeastern.edu | linkedin.com/in/harshitb1611

Education

Northeastern University December 2026
Master of Science in Data Analytics Engineering — GPA: 3.9/4.0 *Boston, MA*

LDRP Institute of Technology and Research May 2024
Bachelor of Engineering in Computer Engineering — GPA: 3.3/4.0 *Gandhinagar, India*

Experience

AI Engineer Co-op Jan 2026 – Present
Staples Inc. *Framingham, MA*

- Fine-tuned and quantized Gemma-series small language models (Unsloth, 4-bit LoRA) and authored custom inference kernels for low-precision execution to replace an LLM planner/router; built a 1,551-row training set, lifting routing accuracy from 62% to 68% while cutting hot-path latency and token cost.
- Profiled the request hot path against a hard latency budget in C++/Python: shipped a Cypher graph-walk scope extractor on Neo4j at p50 110 ms versus a 300 ms budget, and replaced substring entity matching with catalog-grounded LLM extraction to eliminate corrupted retrieval inputs.
- Built and shipped components of a production multi-agent product discovery pipeline (Planner → Discovery → Respond) on LangGraph and Databricks, powering a live B2C conversational shopping assistant over a 200K+ product catalog.
- Co-developed a three-tier conversational memory subsystem on Postgres/pgvector and drove eval-driven delivery via MLflow tracing, LLM judges, and a Prompt Registry behind concrete metric gates.

ML Engineer Intern Jan 2024 – Jul 2024
Pixeltechnologies *Remote*

- Architected Python-based prediction pipelines using scikit-learn and XGBoost to forecast ad engagement on 100K+ events; enhanced campaign ROI by 22%.
- Engineered SQL-driven ETL and Airflow DAGs to process high-velocity A/B test logs; cut model retraining latency by 40%.

Data Analyst Intern Jul 2023 – Jan 2024
GROWITUP *Remote*

- Pioneered a vectorized search system using OpenAI embeddings and FastAPI, achieving 96% faster retrieval across 5K+ documents.
- Modeled user sessions with Snowflake and PostgreSQL to detect behavioral anomalies and surface data drift patterns across engagement funnels.

Projects

Mnemosyne: Agent Memory Engine | *Python, PostgreSQL, pgvector, SQLAlchemy* May 2026 – Present

- Architected a 5-stage async memory pipeline (extraction → embedding → episodic summarization → consolidation → decay) transforming raw agent conversations into structured long-term memory; shipped 10K LOC hardened by 797 tests.
- Engineered a hybrid extraction engine (regex/NER + selective LLM escalation via a zero-cost two-signal routing gate) sending only 20% of sessions to an LLM, cutting blended cost to $0.0004/session.
- Implemented multi-signal scored retrieval (relevance, recency decay, importance, frequency) fused with full-text search via Reciprocal Rank Fusion; pgvector HNSW architected to 50M+ records.

FrontShiftAI: Multi-Tenant Agentic Platform | *FastAPI, Airflow, ChromaDB, Modal, LiveKit, React* Sept 2025

- Designed a Unified Agent Router in FastAPI using regex-based intent detection with LLM fallback to intelligently route requests between deterministic SQL agents and probabilistic LLaMA-3 RAG pipelines.

Technical Skills

Inference & Systems: C++, Python, vLLM, KV-cache & prefill/decode serving, custom inference kernels, low-precision execution, LoRA fine-tuning & quantization (Unsloth, 4-bit), latency/throughput profiling, Linux
Serving & MLOps: Kubernetes, Docker, FastAPI, Prometheus, MLflow, Airflow, CI/CD, SQLAlchemy
AI/ML & LLM: PyTorch, TensorFlow, Hugging Face, Scikit-learn, XGBoost, LangChain, LangGraph, RAG, Multi-Agent Systems, Tool-Calling, OpenAI API
Languages, Retrieval & Cloud: SQL, TypeScript, Cypher; PGVector, Neo4j, ChromaDB, Hybrid Search (BM25 + Vector), RRF; Databricks, GCP, AWS S3, Snowflake