

Shabad Vaswani

New York, New York, United States | shabad.vaswani@nyu.edu | (347) 651-6352 | github.com/shabadvaswani | linkedin.com/in/shabadvaswani | shabad.tech

PROFESSIONAL SUMMARY

Generative AI & Machine Learning Engineer with 3+ years of experience building and deploying scalable AI and machine learning solutions. Skilled in generative AI, LLMs, NLP, deep learning, data engineering, cloud platforms, MLOps, model serving, and distributed systems. Experienced in developing intelligent workflows, production-ready models, data pipelines, APIs, and evaluation frameworks. Strong programming and problem-solving abilities with a focus on scalable architecture, system reliability, performance optimization, automation, and delivering measurable business outcomes.

SKILLS

Programming Languages: Python, Go

Frameworks & Libraries: PyTorch, Transformers, FastAPI, Ethers.js, MetaMask SDK

Databases: Azure Data Lake Storage, Azure Database for PostgreSQL, time-series database

Tools & Software: Docker, Docker Compose, Kubernetes, MLflow, GitHub Actions, Azure Databricks, Braintrust, Datadog, Amazon CloudWatch, LLM Proxy, JSON validation, REST APIs, Azure Machine Learning

Cloud Platforms: AWS, Microsoft Azure, Azure Machine Learning, Azure Databricks, Azure Kubernetes Service (AKS)

Soft Skills: Programming, Problem-solving, Collaboration

WORK EXPERIENCE

Generative AI Engineer – AI Agents

Aug 2025 -

Zendesk

- Developed generative AI agents using LLMs, RAG, MCP, prompt engineering, tool calling, and context management to handle customer requests, improving verified resolution rates by 12% across support conversations.
- Built AI knowledge workflows using embeddings, vector search, Elasticsearch, reranking, and retrieval pipelines, grounding AI-generated responses in trusted support content and improving response relevance by 11% across customer interactions.
- Integrated agentic AI workflows with generative procedures, API actions, MCP tools, LLM routing, and human handoffs, enabling 50+ AI agents to manage multi-step customer service requests across digital support channels.
- Created LLM-powered conversation workflows using prompt templates, structured outputs, conversation memory, guardrails, and fallback handling, reducing unsupported or incomplete AI responses by 10% while improving consistency across customer interactions.
- Implemented GenAI inference services using Python, PyTorch, BentoML, Kubernetes, Docker, AWS, and LLM Proxy, supporting model deployment, routing, scaling, version control, and reliable production responses for AI agents.
- Evaluated generative AI quality using Braintrust, MLflow, Datadog, CloudWatch, and custom evaluation suites, tracking verified resolutions, contained resolutions, relevance, latency, safety, cost, and regression results across AI workflows.

Machine Learning Engineer – Data Science & Engineering

Mar 2022 - May 2024

S&P Global

- Engineered a financial entity resolution platform to standardize company identities across filings, datasets, and market records, improving entity matching precision by 18% across high-volume financial data onboarding workflows globally.
- Developed hybrid NLP models using PyTorch, Transformers, embeddings, semantic similarity, and gradient boosting to rank candidate entities, processing 2.4M records monthly and improving top-5 retrieval accuracy to 94% across validation datasets globally.
- Built distributed ML pipelines using Python, PySpark, Spark, Azure Data Lake Storage, and Azure Databricks for preprocessing, normalization, deduplication, and feature engineering, reducing data preparation time by 35% across 15M monthly records.
- Designed REST-based inference services using Docker, Azure Kubernetes Service (AKS), Azure Database for PostgreSQL, achieving p95 latency of 120ms while enabling confidence scoring for production entity matching.
- Implemented evaluation frameworks using precision, recall, F1-score, ROC-AUC, ranking metrics, and error analysis, increasing F1-score by 12% through threshold tuning, candidate retrieval optimization, and model calibration across financial entity matching tasks.
- Operationalized ML lifecycle workflows covering training, inference, versioning, monitoring, drift detection, and retraining using Azure Machine Learning, collaborating with engineering teams to maintain reliable entity intelligence services across 70M-company-scale reference datasets.

PROJECTS

STEWARD — Agentic Portfolio Manager

- Engineered autonomous quantitative trading engine using four-stage funnel architecture to dynamically screen equities and allocate multi-asset portfolios through LLM agents, integrating GraphRAG knowledge graphs, vector search, Go concurrency, JSON validation, and Docker deployment.

MedAI — Scalable Chest X-Ray Diagnosis

- Deployed containerized FastAPI inference services on Kubernetes, enabling pod scaling for variable chest X-ray workloads, while tracking experiments and model versions with MLflow and automating validation and model redeployment through GitHub Actions pipelines.

AEON — Predictive Wellness Engine

- Co-led design of scalable multi-source pipeline ingesting and synchronizing 40,971 minute-level records across 30 features, automating binary FIT parsing and temporal alignment to merge heterogeneous sources into a unified time-series database for analysis.

Rarity.AI — Generative NFT Platform

- Containerized frontend and backend services with Docker Compose, refined prompts using Groq LLMs, generated images through Gemini API, and integrated Ethers.js with MetaMask SDK to enable secure wallet connections and blockchain transaction signing.

EDUCATION

New York University

Master of Science, Computer Engineering

Ganpat University

Bachelor of Technology, Computer Engineering