

KUSHAL GADDAMWAR

Agentic AI Engineer | Generative AI Engineer | LLMs, RAG, and Multi-Agent Systems

Boston, MA | +1 (857) 328-4611 | kushal7887pd@gmail.com

linkedin.com/in/kushal-gaddamwar | github.com/Kushal9889 | kushal-portfolio-223.netlify.app

Summary

Agentic AI Engineer and M.S. Computer Science candidate at Boston University. Shipped a production Agentic Retrieval Augmented Generation platform on Microsoft Azure for an enterprise client – multi-agent orchestration, hybrid retrieval, knowledge graph – owned from ingestion to deployment. Specializes in Large Language Model (LLM) applications and Generative AI (GenAI) systems engineered for grounding, observability, and low latency. Reported a LangChain SDK defect a maintainer merged. IEEE-published author.

Technical Skills

Languages: Python, SQL, TypeScript, JavaScript, C++

LLM and Agentic AI: AI Agents, Multi-Agent Orchestration, Prompt Engineering, Context Engineering, Tool Calling, Function Calling, ReAct, Model Context Protocol (MCP), OpenAI API, Azure OpenAI

Frameworks: LangChain, LangGraph, LangSmith, LlamaIndex, NVIDIA NIM

RAG and Retrieval: Retrieval-Augmented Generation (RAG), Agentic RAG, Vector Databases, Embeddings, Semantic Search, Hybrid Search, Reranking, pgvector, ChromaDB, Azure AI Search, NV-Embed

Evaluation and Reliability: LLM Evaluation, LLM-as-a-Judge, Guardrails, Grounding, Hallucination Reduction, PII Redaction, LLM Observability, Tracing

Backend and MLOps: FastAPI, RESTful APIs, Microservices, Async Python, Pydantic, Structured Outputs, Docker, Git, GitHub Actions, CI/CD, MLOps, Model Deployment, Low-Latency Inference, Server-Sent Events (SSE)

Databases: PostgreSQL, MySQL, MongoDB, Redis; Graph Databases: Neo4j, Cosmos DB

Cloud: Amazon Web Services (AWS): EC2, S3, Lambda, API Gateway; Microsoft Azure; Google Cloud Platform (GCP)

Frontend: React, Redux, Next.js

Foundations: Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), Deep Learning, Data Structures, Algorithms

Experience

Boston University, Questrom Computational Lab

AI Engineer (Graduate Researcher)

Boston, MA

May 2026 – Present

- Architected an enterprise document-intelligence RAG assistant for an enterprise consulting client on Azure (Azure OpenAI GPT-4o, Azure AI Search), exposing a LangGraph agent with 14 tools for document question answering, comparison, and template-driven generation.
- Engineered hybrid retrieval fusing BM25 keyword search with vector search, adding LLM-based query rewriting and Cohere re-ranking to sharpen answer relevance.
- Ran LLM-as-a-Judge evaluations on assistant responses to measure hallucination and retrieval quality.
- Implemented PII and entity redaction with guardrails across ingested documents, lowering data-leakage risk on sensitive client data.
- Designed header-aware chunking and ingestion (Markdown-header plus recursive splitting) with SHA-256 deduplication, text embeddings, and Azure Blob Storage metadata, indexing the full corpus into a searchable vector index.
- Shipped streaming FastAPI endpoints (SSE) with a React single-page app; modeled a Cosmos DB (Gremlin) knowledge graph linking clients, projects, and technologies.

IMG Systems

Software Engineering Intern

Remote

Aug 2024 – Apr 2025

- Built a document-parsing pipeline (Apache Tika) that raised extraction accuracy 20% and processed 5,000+ candidate profiles monthly, cutting recruiter screening time 15%.
- Enforced structured-output validation with Pydantic against a JSON Schema, reaching 95% schema accuracy and reducing manual review across a 6-person team.
- Containerized Python microservices backed by PostgreSQL and Redis with Docker, trimming REST API latency 25%; automated continuous integration and continuous deployment (CI/CD) via GitHub Actions for zero-downtime releases.

Growaza Pvt. Ltd.

Associate Software Engineer Intern

India

Jan 2024 – Jul 2024

- Optimized API response time 30% via in-memory caching and asynchronous request handling, lifting engagement 22%

- for 1,000+ daily active users.
- Launched a MySQL inventory dashboard tracking 2,000+ SKUs.
- Secured REST endpoints with JSON Web Tokens (JWT) and role-based access control (RBAC); deployed on AWS EC2 and S3.

Publications

- IEEE ICAICCIT 2024 (Author) – [Deep Learning for Contextual Bug Detection](#).
- IGI Global 2024 (Co-Author) – [Cyber-Physical Systems](#).

Projects

LangChain deepagents — Open Source Contribution

Open Source, Python, AI Agents, Async I/O

Jul 2026

- Diagnosed a silent reliability defect in LangChain’s deepagents SDK, an agent framework used across the open-source AI ecosystem, where an unreachable storage backend was reported to the agent as healthy and empty, causing agents to act confidently on missing data instead of surfacing the failure.
- Established it as a recurring class of error-handling defect rather than an isolated bug by tracing the same fault pattern through a previously fixed component, and delivered a minimal reproduction covering both synchronous and asynchronous execution paths so maintainers could confirm it without further investigation.
- Fix reviewed and merged by a LangChain maintainer 57 hours after the report, with the reproduction credited in the [merged pull request](#); the correction now ships to every downstream user of the SDK.

Agentic Portfolio — Production Retrieval Agent

LangGraph, TypeScript, Next.js, RAG, SSE

Aug 2026

- Designed and deployed a live multi-node LangGraph agent that answers recruiter questions from a grounded corpus, streaming responses over Server-Sent Events (SSE) and exposing retrieval sources, per-node latency, and token cost with every answer so any claim can be verified as it is made.
- Held median answer latency at 1.3s by fusing BM25 keyword search with pre-computed embeddings through reciprocal rank fusion, short-circuiting the embedding round trip when lexical scores are decisive.
- Gated every release on a 16-case LLM evaluation suite covering grounding, prompt-injection refusal, and answer consistency, with automatic provider failover across NVIDIA NIM, OpenAI, and Azure OpenAI to keep the service available ([source](#)).

Education

Boston University

M.S. Computer Science

Boston, MA

Sep 2025 – Dec 2026 (Expected)

IIT-DM Jabalpur

B.Tech, Computer Science and Engineering

India

Oct 2020 – Jun 2024

Certifications

- [NVIDIA-Certified Professional: Agentic AI \(NCP-AAI\)](#) (2026)
- [AWS Cloud Technical Essentials](#) (2026)
- IBM RAG and Agentic AI Professional Certificate
- [Google Cloud Fundamentals](#) (2025)

Achievements

- JEE Mains: 99.1 percentile of 1.2M candidates (top 0.9% nationally).
- CodeChef global rank 64 (Feb 2022); 3rd place, Code Rumble 2023.
- Co-founded BITBYTE, a 70+ member coding community; mentees reached Google Kickstart top-200, GSoC, and 3 internships.
- Advisory Head, Tarang Fest 2023: led programming across 13 colleges, 250+ students; raised \$12,000 from 25+ sponsors.