KOUSHIK VASA

Generative AI Engineer | Agentic AI & LLM Systems

Hauppauge, NY • +1 (571) 420-2558 • vasakoushik@gmail.com • [LinkedIn] • [GitHub] • [Portfolio]

PROFESSIONAL SUMMARY

Generative AI Engineer with **5+ years** building production-grade agentic systems end-to-end — from RAG pipelines and multi-agent orchestration to inference optimization. Reduced hallucinations by 25–27% and latency by 40–50% across healthcare and B2B SaaS environments. Most recently served as the sole AI engineer at Bridge AI, owning end-to-end agent architecture, tool orchestration, and system performance using Claude and DSPy.

TECHNICAL SKILLS

Generative AI & LLM Systems: Agentic AI, RAG (Hybrid, Re-Ranking), Embeddings, Semantic Search, Prompt Engineering, Tool Calling, Agent Orchestration

LLM Frameworks & Platforms: LangChain, LangGraph, DSPy, Azure OpenAI, OpenAI API, Anthropic Claude, Hugging Face

ML & NLP: PyTorch, TensorFlow, Scikit-learn, NLP, Classification, Feature Engineering, Model Evaluation

Infrastructure & Cloud: Azure (ML, AI Search, Databricks), AWS (S3, Lambda), Docker, Kubernetes, CI/CD (Azure DevOps, GitHub Actions)

Backend & Data: FastAPI, Flask, Supabase (PostgreSQL), pgvector, RBAC, OAuth

Languages: Python, SQL, TypeScript, Bash

EDUCATION

Master of Science in Computer Science (Machine Learning) — GPA 3.87 *Jan 2024 – Dec 2025*

George Mason University, Fairfax, VA

PROFESSIONAL EXPERIENCE

AI/ML Engineer

Aug 2025 – Present

Bridge AI | USA

- Built Bridgette, an agentic AI assistant for a B2B marketplace, replacing a 10+ step workflow (search → filter → invite → negotiate) with a single conversational interface.
- Designed a multi-step tool-calling agent using Claude (Anthropic API) and DSPy for structured prompt optimization and execution control.
- Designed agent decision framework enabling autonomous execution across retrieval, structured queries, and response generation.
- Architected production dual-agent system (companies vs. experts) supporting workflows such as expert matching, engagement creation, and earnings analysis via natural language.
- Developed 8 production-grade tool schemas with strict runtime validation (Zod), ensuring reliable execution across agent decision and tool-calling layers.
- Improved end-to-end agent task success rate from ~62% to ~84% by identifying and fixing gaps across retrieval, reasoning, and execution layers.
- Reduced hallucinated or incomplete responses by ~25% through validation layers and fallback mechanisms, evaluated across ~500+ test queries.
- Reduced average response time from ~1.2s to sub-400ms via prompt optimization and query-level caching.
- Sole AI engineer — owned end-to-end system design covering agent architecture, tool orchestration, backend APIs, and frontend streaming integration.

AI Engineer

Apr 2021 – Dec 2023

Capgemini | India

- Built document intelligence platform using hybrid RAG (Azure OpenAI + BM25), replacing manual document review for ~120 internal users across multiple teams.
- Improved retrieval accuracy for enterprise document queries by ~32% by combining dense embeddings with BM25 re-ranking across 500K+ unstructured documents.
- Designed LangGraph-based pipelines with clear separation of retrieval, reasoning, and response generation, improving response reliability by ~35%, reducing incomplete and inconsistent outputs in production.
- Developed FastAPI inference services handling 200+ concurrent requests with consistent sub-second response times.
- Introduced confidence-based validation and context filtering that reduced hallucinated or irrelevant responses by ~27% based on internal evaluation benchmarks.
- Cut inference latency by ~40% through prompt compression, caching, and retrieval pre-filtering.
- Containerized services using Docker and implemented CI/CD pipelines, reducing deployment time from hours to ~15 minutes.

Machine Learning Engineer *Aug 2019 – Mar 2021*

CitiusTech Healthcare Technology Pvt. Ltd. | India

- Built Python/SQL pipelines processing 1M+ patient and claims records, reducing data preparation time by ~30% for downstream ML workflows.
- Developed 30-day hospital readmission prediction models using XGBoost and Random Forest, achieving ~82% AUC and enabling early identification of high-risk patients.
- Applied PCA and domain-driven feature engineering on clinical and claims data, improving model F1-score by ~12% and reducing feature dimensionality by ~40%.
- Resolved recurring pipeline failures including data inconsistencies and job timeouts, reducing failure rates by ~35% while consistently meeting SLA timelines.

PROJECTS

MediConnect: AI-Powered Healthcare Matching Platform (2025) | React, Node.js, Gemini, SQLite, Supabase

- Built and deployed AI-powered patient-to-doctor matching platform using 2.8M+ CMS clinician records, enabling real-time symptom-based specialist discovery
- Designed Gemini 2.5 Flash–based recommendation engine with confidence scoring and fallback handling for ambiguous inputs
- Developed multi-factor compatibility scoring system ranking doctors based on symptoms, geolocation, and predicted consultation cost
- Built conversational AI assistant supporting structured intake, image uploads, and contextual diagnostic guidance
- Implemented Node.js + Express APIs (5 endpoints) powering search, ranking, and recommendation workflows
- Designed interactive frontend with OpenStreetMap-based distance calculations and 3D anatomy explorer (Three.js) tied to medical conditions
- Integrated Supabase authentication, user profiles, and emergency location-sharing features

ClearCare: AI-Powered Clinical Data Pipeline *(2023)* | Python, SQL, NLP

- Built end-to-end pipeline ingesting structured EHR and unstructured clinical notes, normalizing data for downstream ML and analytics workflows
- Automated validation and transformation in Python, reducing manual preprocessing effort and recurring data quality issues
- Integrated LLM-based entity extraction to standardize clinical terminology, improving consistency of training datasets

CitationSleuth: RAG-Based Fact Verification System *(2024)* | Python, Neo4j, Embeddings, RAG

- Built dual-layer LLM validation system combining semantic retrieval and Neo4j graph traversal to verify generated claims
- Improved verification precision by linking embedding-based evidence retrieval with graph-based relationship validation
- Developed real-time interface surfacing low-confidence or unsupported outputs before downstream usage

CERTIFICATIONS

- Generative AI LLMs — NVIDIA (Jan 2026)
- Microsoft Fabric Analytics — Microsoft (Jan 2026)
- Agentforce — Salesforce (Dec 2025)
- OCI AI Foundations Associate — Oracle (Oct 2024)
- Generative AI Fundamentals — Databricks (Aug 2024)
- Prompt Engineering for Everyone — IBM (Aug 2024)