

Prakhar Varshney

I BUILD SYSTEMS THAT **THINK TOGETHER.**

New Delhi, India / prakhar@bu.edu / +91 98183 11267

github.com/prkhrv / linkedin.com/in/prkhrv

CURRENT FOCUS

Senior AI systems engineer building production multi-agent platforms, evaluation harnesses, and distributed infrastructure that turns ambiguous human workflows into reliable automation.

OPERATING MODE

Agent architecture • Model orchestration • AI cost optimization • Evals & hallucination handling • Product engineering • AWS serverless systems

01 EXPERIENCE

Sourcy Global

2024 — Present

Senior Software Engineer — Multi-Agent AI Systems

Singapore / Remote

- **Built the agent harness.** Engineered the production runtime behind an autonomous sourcing team: specialized agents, Zod-validated tool contracts, shared workflow state, checkpointed execution, retries, fallbacks, and human escalation across **Gemini, OpenAI, and Anthropic**.
- **Made quality measurable.** Built constraint and schema validation, evidence grounding, hallucination checks, confidence gates, regression evals, and LLM-as-judge scoring—sustaining **95%+ specification accuracy across 1,200+ monthly requests**.
- **Optimized every token.** Reduced AI cost and latency through task-aware model routing, context compaction, token budgets, caching, batching, parallel fan-out, and early exits—reaching **100+ specifications per minute and 70% faster turnaround**.
- **Productionized the intelligence.** Ran replayable, observable workflows on **AWS Lambda, Step Functions, SQS, CDK, Python, TypeScript, and PostgreSQL**, with per-run cost/latency traces and recovery paths powering AutoQuotation, supplier ranking, and product mapping.

Patient First.AI

2022 — 2024

Senior Software Engineer — Healthcare AI

Boston, USA

- **Automated compliance.** Launched **VacTrack**, an AI vaccination-compliance platform built with Flutter, React, NestJS, MongoDB, and AWS, reducing Student Health Services workload by approximately **60%**.
- **Scaled document intelligence.** Architected and horizontally scaled a Python document-AI pipeline to **1,000 documents per minute** using Docker, AWS CDK, Lambda, SQS, SNS, and real-time Socket.io updates.
- **Returned time to physicians.** Built an AI-driven patient engagement platform across SMS, email, phone, and WhatsApp, saving physicians an average of **three hours per day**.

TutorBin

2021

Software Engineer — Platform Systems

India

- **Modernized operations.** Led delivery and containerized deployments of operational dashboards with **Python, Django, Vue, Kubernetes, Docker, and S3**, increasing marketing throughput by **30%**.
- **Accelerated payments.** Built an event-sourced **Kafka** payment service that made client-side payments **65% faster**.
- **Moved work off the request path.** Added Celery, Django Channels, and pyPDF workflows that improved platform speed by **18%** and reduced server cost by **21%**.

OPERATING RANGE

02 / SYSTEMS & SELECTED WORK
START A CONVERSATION →

AGENTIC AI	Multi-agent architecture, agent harnesses, model routing, tool use, RAG, prompt/context engineering, memory and state, human-in-the-loop workflows
AI RELIABILITY	Offline and regression evals, LLM-as-judge, hallucination detection, grounding, confidence gates, schema/constraint validation, cost and token telemetry
DISTRIBUTED SYSTEMS	Serverless workloads, event-driven architecture, queues, microservices, real-time systems, checkpointing, retries, replay and failure recovery
PRODUCT ENGINEERING	React, Next.js, TypeScript, Node.js, NestJS, Python, Django, Flask, Flutter, Zod, Tailwind CSS
CLOUD & DATA	AWS CDK, Lambda, Step Functions, SQS, SNS, PostgreSQL, MongoDB, DynamoDB, Docker, Kubernetes, GCP, Azure
LANGUAGES	Python, JavaScript / TypeScript, Java, C / C++, Dart, SQL, Bash, HTML / CSS

03 EDUCATION

Boston University	2021 — 2023
Master of Science in Computer Science	Boston, USA
Noida Institute of Engineering and Technology	2016 — 2020
Bachelor of Technology in Computer Science and Engineering	Delhi NCR, India

04 SELECTED PROJECTS

API-CLI →	Open-source Node.js CLI that generates a working REST API foundation in one command.	NODE / NPM
Meet.OZ →	AI platform for analyzing participant emotion and engagement in online meetings and classes.	REACT / FLASK
Holistic →	Mobile mental-health and wellness companion with personalized profiles and virtual coaching.	FLUTTER / AWS

05 RESEARCH & RECOGNITION

BU Questrom Research	Built a Python and BigQuery pipeline for InnoVAE, transforming unstructured patent text into an interpretable spatial representation of innovation and developing a React and Flask model-performance dashboard.
BostonHacks Winner	Won BostonHacks 2021 by building a mobile application for tracking and supporting mental health.
NCAE CyberGames	Ranked 3rd in the 2022 Northeast Conference competition.

FROM INTERFACES → DISTRIBUTED SYSTEMS → COORDINATED INTELLIGENCE

Every layer matters when the system has to earn trust. PRAKHAR@BU.EDU →