Muhammad Hassan Baig

Boston, MA | hassanbaigy3@gmail.com | linkedin.com/in/hassanbaigy | github.com/hassanbaigy

SUMMARY

AI engineer with 4+ years in Python and TypeScript and 2.5 years shipping production LLM and agent systems on LangChain and LangGraph. As founding engineer on a 1,000+ user agent platform, turned customer workflows into specs, built the agents, curated training data, and ran evals and deployments.

TECHNICAL SKILLS

Agent engineering: LangChain, LangGraph, Deep Agents, LangSmith (tracing, datasets, evals), multi-agent orchestration, short- and long-term memory, tool calling, MCP servers, prompt-injection fencing, streaming chat UIs
Retrieval and vector search: RAG, OpenSearch k-NN, Elasticsearch, Pinecone, pgvector, HNSW index tuning, hybrid search, reranking
Applied AI: dataset curation, LoRA fine-tuning, evals (precision/recall/F1, LLM-as-judge), trace mining, model selection
Engineering: Python, FastAPI, PyTorch, SQL, PostgreSQL, Redis, Temporal, Docker, Kubernetes, AWS, TypeScript, React

PROFESSIONAL EXPERIENCE

graph8, Remote, United States — Mar 2025 – Apr 2026
Artificial Intelligence Engineer, founding engineer (Contract)

- Hired by Mohtashim, one of Asia's top LangChain open-source contributors; built graph8's LangChain, LangGraph, and Deep Agents stack under his guidance, with tool calling, LangSmith tracing, and custom MCP servers.
- Rebuilt a Ruby prototype into a generative-AI outreach platform on FastAPI, Next.js, and LangGraph: 1,000+ active users and 2M+ emails per month, with RAG analytics agents, draft generation, and a web-chat agent.
- Built the AI Inbox end to end: LLM intent tagging with a reply-disposition decision tree, noise filtering, and opt-out and interest verdicts at ingestion, with prospect messages fenced as untrusted content.
- Curated labeled datasets from production replies and LangSmith traces, deduplicating, filtering noise, and hand-labeling edge cases, to train and LoRA fine-tune the spam and intent classifiers; scored each version on precision, recall, and F1 against a held-out set before rollout.
- Mined LangSmith traces for failed tool calls and misclassified replies and added them to the eval sets, so each fix became a regression test.
- Cut model inference costs about 50% against third-party APIs by self-hosting Llama and Mistral plus the spam and intent classifiers on Kubernetes with autoscaling and monitored rollouts.
- Designed the deliverability engine on Temporal: durable workflows that pace sending, run warm-up, and throttle on bounce rate across 10k+ mailboxes.
- Turned customer feedback into specs and the roadmap as founding engineer; made the architecture calls and mentored 4 developers through code review.

Northeastern University, Roux Institute, Portland, ME — Sep 2026 – Present
Research Assistant, Cloud AI Specialist (part-time)

- Build and maintain the cloud infrastructure and AI data pipelines for a faculty-led research project.

SageTeck, Islamabad, Pakistan — Nov 2023 – Nov 2025
Software Engineer (full-time Mar 2024 – Mar 2025; part-time advisory otherwise)

- Architected a multi-tenant SaaS platform in React Native, Laravel, and FastAPI, adding LangChain RAG pipelines over Pinecone vector search to generate content from each tenant's own data.
- Raised app performance 40% and crash-free sessions 25% by profiling slow screens and lazy-loading code.

Dubizzle Labs, Lahore, Pakistan — Feb 2023 – Mar 2024
Associate Software Engineer

- Shipped listing and search features in Next.js and TypeScript for Profolio, a real-estate platform with hundreds of thousands of monthly users; added Elasticsearch search and Sentry monitoring.

Tintash, Lahore, Pakistan — Aug 2021 – Jan 2023
Associate Software Engineer

- Built client, trainer, and admin portals in React and Next.js for a personal-training company with 3,500+ clients; led client stand-ups to demo work and gather requirements.

PROJECTS

claude-agent-team, open-source npm package (creator and maintainer) — Jan 2026

- Wrote a CLI on the Claude Agent SDK that installs a lead coordinator and 10 specialist subagents with a single-writer shared-memory protocol for long-term state; a 13-engineer team made 198 commits to its shared memory in 14 days.

EDUCATION

Northeastern University, Roux Institute, Portland, ME — Apr 2026 – Expected Jun 2028
Master of Professional Studies in Analytics, Applied Machine Learning Concentration

GIK Institute of Engineering Sciences and Technology, Topi, Pakistan — Aug 2018 – Jul 2022
Bachelor of Science in Computer Science