

Jaspreet Singh

jsm9323@gmail.com • (+91) 8790547756 • linkedin.com/in/jaspreetm • github.com/jp111 • Bangalore, India

Applied AI Engineer with 8+ years building production ML and large-scale distributed systems. Ships agentic LLM applications end to end — multi-agent orchestration on LangGraph, RAG grounded in knowledge graphs, human-in-the-loop guardrails, and evaluation harnesses — on AWS streaming and batch pipelines at 10M+ records daily. Research roots in neural retrieval and ranking.

TECHNICAL SKILLS

- **Generative AI:** LLMs, Retrieval-Augmented Generation (RAG), GraphRAG, Multi-Agent Orchestration, LangGraph, Prompt Engineering, Knowledge Graphs, Agent Memory, Human-in-the-Loop (HITL), Guardrails, Evals, Transformers/BERT, PyTorch
- **Languages & AI Tooling:** Python, Java, Go, C++, SQL, JavaScript | Claude Code, Cursor, GitHub Copilot
- **Data & Cloud:** AWS (Kinesis, S3, Lambda), Kafka, PySpark, Parquet, Redis, Elasticsearch, Docker, Kubernetes, Terraform

PROFESSIONAL EXPERIENCE

- **Apple** Bangalore, India
Applied AI Engineer Oct 2025 - Present
 - Designed a production AI SRE platform for autonomous incident investigation on **LangGraph** multi-agent workflows — stateful graphs with tool nodes, checkpointing, and retries — scaling RCA across **10k+ daily alerts** and cutting **MTTR 30%**.
 - Built the **RAG** retrieval layer over a domain-specific **knowledge graph**, runbooks, prior postmortems, and live telemetry, with persistent long-term memory across incidents, so every generated root cause is grounded in cited evidence rather than inferred.
 - Shipped **human-in-the-loop** approval gates and guardrails on all consequential remediation actions, capturing engineer accept/reject decisions as labeled **evaluation signal** to measure agent precision, robustness, and business impact over time.
 - Led development of the execution engine, wiring real-time telemetry into graph-driven remediation workflows in Python on AWS.
- **Block Scholes** London, Remote
Senior Software Engineer Sep 2024 - Sep 2025
 - Architected a high-throughput streaming and batch pipeline on **AWS Kinesis** and S3 processing **10M+ daily records**, with end-to-end monitoring, logging, and metrics, and raised test coverage 30% to harden it for production.
 - **Cut storage costs 40%** via **PySpark** Parquet compaction and partition tuning across the S3 data lake.
- **TikTok Live (ByteDance)** Singapore
Senior Software Engineer Dec 2022 - Dec 2023
 - Led backend development of creator observability pipelines, serving real-time analytics dashboards at **350K+ QPS**.
 - Built an Anti-Money Laundering (AML) risk-detection system over streaming behavioral signals across **7M+ live rooms daily**.
 - Migrated core storage to a Redis architecture, **reducing read latency 65%** for high-concurrency queries.
- **Goldman Sachs** Bangalore, India
Associate / SDE-II (L4) Jan 2018 - Nov 2022
 - Led an observability platform indexing **11M+ queries/day** across Elasticsearch, powering firmwide search and alerting.
 - Built fault-tolerant AWS microservices and a batch orchestration layer for **10k+ scheduled jobs**, cutting incidents to near zero.

RESEARCH EXPERIENCE

- **Thesis: Automatic Information Retrieval for Short Documents — IIIT Hyderabad** Hyderabad, India
Advisor: Dr. Vikram Pudi May 2016 - Aug 2018
 - Built an NLP information retrieval system saving **31% screening time at 100% recall** across **117K+ PubMed** documents.
 - Developed NITBUG, a **BERT**-based dense retrieval model using triplet learning and inter-document context, achieving **10–15% higher recall and 61–65% Recall@1** across Mozilla, Eclipse, and NetBeans corpora.

EDUCATION

- **International Institute of Information Technology, Hyderabad** Telangana, India
M.S. by Research, Computer Science and Engineering | GPA: 8.6/10.0 Aug 2016 - Apr 2018
- **International Institute of Information Technology, Hyderabad** Telangana, India
B.Tech (Honours), Computer Science and Engineering | GPA: 8.6/10.0 Aug 2012 - Apr 2016

ACHIEVEMENTS

- **Competitive:** 3rd place, Google Cloud Developer Challenge (South Asia) | Rank 15, Google APAC 2017 | ACM-ICPC Asia Regionalist | Dean's Merit List (Top 5%), IIIT Hyderabad.