Simran Malik

San Diego, CA | +1 (858) 214-4589 | malik.simran4@gmail.com | linkedin.com/in/simranmalik4 | github.com/simran-malik

PROFESSIONAL SUMMARY

AI Engineer with **4+ years at Google** shipping software and distributed systems at **12M+ daily orders**, now building **AI agents**: at Brain Corp, designed a **multi-agent retrieval system** over a knowledge graph with verification by an evaluator agent and deterministic safeguards (**patent in preparation**). **NeurIPS 2026 co-author** on evaluating AI answer sources. AI-native, with strong problem-solving skills; uses **Claude Code** daily. **Twice recognized for self-driven ownership at Google.**

EDUCATION

Master of Science in Computer Science (AI), University of California San Diego — GPA: **3.96/4.0** | Expected Dec 2026

Bachelor of Engineering in Computer Science, BITS Pilani — GPA: **9.51/10** | Aug 2021

WORK EXPERIENCE

Brain Corp – Platform Intern (ML & SWE) — San Diego, CA | **Jun 2026 – Sep 2026**

- Developed a **multi-agent retrieval system (patent in preparation)** whose LLM agents traverse a knowledge graph to answer store questions, with verification by an evaluator agent and **6+ deterministic safeguards** for audit trails.
- Constructed a novel **knowledge graph** of stores up to **250K+ sq ft** from robot images and vector embeddings to ground agents; computed spatial relations from geometry instead of an LLM, saving **~64M input tokens**.
- Architected a customer-facing **task-time estimator** for **2500+ robots** across database, cloud, robot and mobile app, blending historical and new runs weighted by observation count to stay accurate for new or rare tasks.

Google – Software Engineer III — Bangalore, India | **Jul 2021 – Aug 2025**

- Released user funds for **300K+ orders** stuck for months (**99% reduction**) by diagnosing execution-state inconsistencies in long-running payment workflows, then shipped automated monitoring and observability to prevent recurrence.
- Protected **12M+ daily orders** by working directly with enterprise partners to migrate APIs to higher-availability systems, cutting failures from **2.80% to 1.28%** (**tech-lead bonus**); built a reusable partner-onboarding framework.
- Prevented **$4M+ revenue loss** by designing a caching-based system that kept bill-payment and recharge workflows reliable during bank-system failures, cutting throttled orders from **46K+ to zero** and repeated retries by **20%**.
- Owned an invoicing platform for a venture with a **$25M+ revenue** projection from concept to production; built third-party integrations and led communication across Business, Tax, and UX stakeholders (**Director's Spot Bonus**).

Apple – Intern - Ad Platform (Software Engineering) — Hyderabad, India | **Jan 2021 – Jul 2021**

- Designed and built a **data management portal** for the Apple Ads platform, saving **10+ engineering hours** per week.

PUBLICATIONS

SourceBench: Can AI Answers Reference Quality Web Sources? | ***NeurIPS 2026**, Evaluations & Datasets Track* — Accepted

H. Jin*, S. Liu*, Y. Li*, **S. Malik**, Y. Zhang (*equal contribution)

- Built the **evaluation pipeline** measuring the **provenance** and quality of sources cited by **12 AI systems** on **8 metrics** across **100 questions**; proposed the **research direction** and human-labeled ground truth.
- Led the Google vs. AI-search study, showing top AI systems cite sources largely absent from Google's top results (**16% overlap** for GPT-5 vs. **55%** for Tavily); caught a **prompt-bias bug** that would have skewed every system's results.

PROJECTS

Natural-Language Q&A over Robot-Built 3D Maps | *UCSD Existential Robotics Lab* — Sep 2026 – Present

- Raised F1 up to **3.5x** on initial test queries by redesigning prompts for an **agentic system** that answers natural-language questions over a robot-built multimodal map; built the **evaluation harness** (precision, recall, F1, AP).

RAG-powered Q&A Chatbot for Technical Documentation | *LangChain, FAISS, HuggingFace* — Apr – May 2026

- Ranked **3rd of 141 teams** (F1@5 **94.7%**, **100% faithfulness**) with a multi-stage RAG pipeline over FAISS retrieval; two-stage reranking lifted validation F1@5 from **65% to 93%**.
- Achieved the **#1 retrieval scores** (precision, recall, F1) of all 141 teams by noticing the gold labels favored a few key sources and designing a precision-first funnel: cross-encoder reranking followed by LLM filtering.

Fine-Tuning Small LLMs to Map Questions to Database Schema | *HuggingFace PEFT, Qwen2.5-Coder* — May – Jun 2026

- Improved schema linking **6.5x (9.7% to 62.7%)** over 13 experiments with per-stage LoRA adapters on Qwen2.5-Coder.

TECHNICAL SKILLS

AI / Agents: Multi-Agent Orchestration, Agentic Workflows, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Context Engineering, Knowledge Graphs, LLM Evaluation & Verification, Fine-tuning (LoRA/PEFT), PyTorch

Backend & Cloud: Backend Services, Distributed Systems, Microservices, REST & gRPC APIs, Caching, Monitoring & Observability, Database Design, PostgreSQL, MySQL, GCP, Docker, Kubernetes, Infrastructure as Code (IaC) with Pulumi, Git

Programming Languages & AI Tools: Python, Java, Go, TypeScript, C++, SQL, LangChain, Claude Code, Cursor, Codex