

Rayhan Patel

Software Engineer, AI Agents | Production Agents, Agent Memory, LLM Evaluation

rayhanbp@umd.edu | github.com/Rayhanpatel | linkedin.com/in/rayhan-patel-cs | chat.rayhanpatel.com

San Francisco Bay Area, CA | M.S. May 2027 | Available June 2027 | Open to relocation nationwide

Experience

3E, AI Platform team

Jun 2026 – Aug 2026

AI Engineer Intern | agent memory & LLM evaluation for a production compliance-AI platform | Bethesda, MD

- Implemented the core of a multi-tenant **agent-memory MCP service** on AWS Lambda and DynamoDB: six operations, per-tenant access that denies by default, versioned writes; built its deterministic spec-conformance harness (~125 tests).
- Benchmarked three open-source agent-memory systems (Hindsight, Cognee, Zep/Graphiti) on **LoCoMo** and **LongMemEval** under a locked protocol: pinned prompts, sha-verified question sets.
- Calibrated the **LLM judge against blinded human graders**. The scoring rubric, not the judge model, turned out to be the accuracy lever.

Euler AI, two-product AI startup

Mar 2025 – Jul 2025

Founding ML/Software Engineer (Full-Time, Remote)

- Built a **conversational shopping agent** as FastAPI microservices (AsyncIO): intent routing, a reranker, PII guardrails.
- Shipped an LLM-as-a-judge (G-Eval) batch pipeline that measured hallucination across the product.

Projects

Career Agent Platform, live production agent

chat.rayhanpatel.com | github.com/Rayhanpatel/AI-Resume-Agent

- Deployed a production AI agent on Gemini: **10 backend modules** with graceful degradation, plus five automated LLM-as-judge evaluators (Langfuse, 100% sampling) scoring every answer.

HVAC Copilot, on-device voice + vision agent (4-person hackathon)

github.com/Rayhanpatel/Ycombinator-Cactus-Deepmind

- Designed a six-tool function-calling surface on Gemma 4 E4B, fully on-device (Cactus SDK), with the on-device/online boundary enforced in the dispatcher, not just the prompt.
- Profiled it end to end: **TTFT 217 ms bare-model against 3.9–4.5 s** once ~935 prefill tokens of system prompt and tool schemas loaded. Root-caused to CPU prefill, not the model.

Multi-Objective LLM Inference Tuning, 3-person UMD study; owned one of the five methods compared

- Designed the agent loop, benchmarked on **Slurm across four GPU allocations**: it proposes one vLLM serving config per trial across nine knobs, optimising throughput, TTFT and TPOT. Every config passes schema repair before it reaches a GPU.

Other projects

- Wrote a Top-K k-NN kernel in CSL for the Cerebras WSE-2 (take-home; code on request): all 6 grader cases pass, **2.68× fewer cycles** than the first correct version; a radix-sort variant was measured and rejected at +24%
- Built *MedBill*, a medical-bill auditor, solo with a parallel **Claude Code** workflow, and *RegSentinel*, a compliance agent on the Claude Agent SDK and MCP github.com/Rayhanpatel/medbill

Publications

EvoRank: LLM-Guided Evolution of Multi-Objective Learning-to-Rank Pipelines

RecSys 2026 workshop

First author, equal contribution: LLM-guided evolution of LTR pipelines gated by a pre-spend headroom check and a held-out transfer audit; beat a tuned LambdaMART on relevance in all three runs (best: **+0.0057 NDCG@10**, $p < 0.0001$, 59,902 held-out queries) and on every objective in the best run; **20th of 340** on the Kaggle hidden test set from a one-shot late submission. arxiv.org/abs/2609.22196

Building Domain-Specific LLMs Faithful to the Islamic Worldview

NeurIPS 2023 workshop

Co-author: tested whether prompting, RAG over 54K Hadith or GPT-3.5 fine-tuning makes LLMs answer Islamic questions faithfully to scholar answers; built the BERTScore evaluation and presented the paper. arxiv.org/abs/2312.06652

Technical Skills

Languages: Python (AsyncIO), TypeScript, JavaScript, SQL, C, Bash

AI & LLM Systems: LLM Agents, Agent Memory, LLM Evaluation & Benchmarking, LLM-as-a-Judge (G-Eval), RAG, MCP, Function Calling, Guardrails, On-Device LLM, NLP, Information Retrieval

Frameworks & Tooling: Claude Code, Claude Agent SDK, FastAPI, pytest, vLLM, Node.js, Mem0/EmbedChain (OSS)

Systems & Infrastructure: Numba CUDA kernels, distributed & parallel computing (Slurm multi-GPU), Linux, Docker, AWS (Lambda, DynamoDB), GCP (Vertex AI), PostgreSQL, Git, GitHub Actions; AWS Certified AI Practitioner, 2026

Education

University of Maryland, College Park

Aug 2025 – May 2027

M.S., Applied Machine Learning | Algorithms & Data Structures, Computing Systems for ML, Optimization

B.S. Abdur Rahman Crescent Institute of Science & Technology

Aug 2019 – May 2024

B.Tech., Computer Science & Engineering | Chennai, India