Rayhan Patel

rayhanbp@umd.edu · +1 (650) 521-4222 · LinkedIn · GitHub · chat.rayhanpatel.com
College Park, MD · F-1 CPT eligible Summer 2026 (zero cost to employer) · Open to relocation

Education

University of Maryland, College Park — Aug 2025 – May 2027
M.S., Applied Machine Learning — Deep Learning, NLP, Computer Vision, Optimization

B.S. Abdur Rahman Crescent Institute of Science & Technology — Aug 2020 – May 2024
B.Tech., Computer Science & Engineering | First Class Honours — Chennai, India

Experience

Euler AI — Mar 2025 – Jul 2025
Founding ML/Software Engineer (Full-Time, Remote)

- Built an **e-commerce conversational AI agent** end-to-end: designed intent classifiers and **query reconstruction** for product queries in natural language, integrated a **reranker for search relevance and personalized discovery**, and implemented PII guardrails with prompt injection prevention.
- Architected orchestration pipeline: **embedding models for semantic product-catalog matching**, knowledge graph with **persistent memory** for multi-turn context, and multi-model ensemble routing via FastAPI microservices.
- Implemented **LLM-as-a-judge evaluation** (G-Eval) measuring coherence, relevance, and faithfulness in batch, with end-to-end agent tracing for throughput, latency, and cost monitoring.
- Contributed to open-source frameworks **Mem0** and **EmbedChain**: added persistent memory for agentic multi-turn workflows.

Publication

Building Domain-Specific LLMs Faithful to the Islamic Worldview — Dec 2023
NeurIPS 2023 MiML Workshop (Patel et al.) · *arxiv.org/abs/2312.06652*

- LLM alignment via **RAG over 54k-document corpus**, prompt engineering, and GPT-3.5 fine-tuning; evaluated with BERTScore + embedding distance — best F1: 0.402.

Projects

AI Resume Chatbot — Jan 2026 · Live · GitHub · Demo Video
FastAPI, Gemini, httpx/HTTP-2, RAG, AsyncIO, LLM-as-a-Judge

- Built a **production AI agent** with 10 backend services (Intent Classifier, Agent, Memory, Rate Limiter, Session, Lead Capture, Job Extractor, Prompt Generator, Tracer) with graceful degradation across every dependency.
- Replaced blocking GenAI SDK with direct **httpx HTTP/2** calls — async connection pooling (20 keepalive, 100 max) and thread-safe TTLCache with deadlock prevention.
- Deployed 5 automated **LLM-as-a-Judge evaluators** (Hallucination, Relevance, Conciseness, Helpfulness, Toxicity) via Langfuse at 100% sampling, zero latency impact.

PathGuard Edge — Feb 2026 · GitHub · Demo Video
Grounding DINO, SAM2, Depth Anything V2, Liquid LFM 2.5 VLM

- On-device VLM (Apple Neural Engine) generates scene-specific prompts dynamically — **zero-shot object understanding** from images, replacing static detection with **multimodal visual search**.
- Real-time pipeline: Grounding DINO + SAM2 + Depth Anything V2 with state machine + telemetry. **$17.5K prize**, UMD × Ironsite Hackathon.

FunctionGemma Hybrid Router — Feb 2026 · GitHub
FunctionGemma-270M, Gemini 2.5 Flash Lite, Cactus SDK

- 3-tier hybrid inference: on-device model with cloud fallback; lexical pre-router achieves 70% on-device ratio. **0.99 F1 at 548ms**. **2nd Place / $5K**, Cactus × Google DeepMind Hackathon.

English2SQL | B.Tech Final Year Project — 2024 · GitHub
LLM, RAG, FastAPI, Docker, PostgreSQL

- NL-to-SQL over relational and star schemas using **RAG + few-shot prompt optimization**; +9% execution accuracy on 120-query eval suite.

Honors & Awards

Cactus × Google DeepMind Global Hackathon — 2nd Place, Lead ($5,000 Gemini API credits) — Feb 2026
UMD × Ironsite Startup Shell Hackathon — Lead, Team PathGuard ($17,500 prize pool) — Feb 2026
Pear VC + OpenAI Hackathon, San Francisco — Finalist — 2024

Technical Skills

GenAI & LLM	Conversational AI, RAG, Agents, Intent Classification, Query Reconstruction, Reranking, Embeddings, Personalization, Guardrails, G-Eval, Prompt Engineering, Function Calling, LangChain
ML Frameworks	PyTorch, TensorFlow, Transformers (HuggingFace), Scikit-learn, OpenCV, XGBoost, Pandas, NumPy
CV & Vision	Grounding DINO, SAM2, Depth Anything V2, VLMs, Multimodal Search, Zero-Shot Detection, ONNX, FP16
Languages	Python (AsyncIO), TypeScript, SQL, C, Bash
Cloud & Backend	GCP, AWS, Docker, FastAPI, GitHub Actions, CI/CD, Nginx, Linux, WebSockets