Harsh Desai

732-890-8061 | desai.harsh.747@gmail.com | LinkedIn | GitHub

Summary

ML engineer with research experience in LLM safety and guardrail systems. Available June to August 2026 before start of M.Eng degree at Cornell. Experienced in distributed training, retrieval systems, and end-to-end ML pipelines.

Education

Cornell University *August 2026 to May 2027*
Master of Engineering in Computer Science *Ithaca, NY*

University of Maryland, College Park (GPA: 3.85/4.00) *August 2023 to May 2026*
Bachelor of Science in Computer Science, Minor in Statistics *College Park, MD*

Relevant Coursework: Data Science, Machine Learning, Natural Language Processing, Advanced Data Structures, Algorithms, Graph Theory, Linear Algebra, Software Engineering, Computer Vision and Deep Learning

Experience

Undergraduate Machine Learning Researcher *September 2025 to May 2026*
UMD Institute for Advanced Computer Studies (UMIACS) *College Park, MD*

- Contributed to **DynaGuard 2**, a dynamic guardrail model to ensure policy adherence for large language models.
- Built a data pipeline to reformat safety-focused datasets (CWEval, RedCode, MT-SEC, ProSec) into the model's expected input schema, yielding **220K+ samples for training**.
- Supported **vERL and Modal** training pipeline by implementing **GRPO and SFT**, resulting in higher F1 scores across held-out safety benchmarks than other current models.
- Experimented with **major red-teaming frameworks** such as NVIDIA's garak and Microsoft's PyRIT, identifying ease of use and applicability, directly informing project's tool choice.

Advanced Data Structures Undergraduate Teaching Assistant *September 2025 to May 2026*
UMD Department of Computer Science *College Park, MD*

- Held weekly office hours for **up to 500 students**, clarifying advanced data structures including B+ Trees, KD-Trees, and Blockchains, with emphasis on time and space analysis.

Backend Engineer / Engineering Administrator *September 2024 to January 2026*
Hack4Impact *Remote*

- Defined and implemented **scalable PostgreSQL database schemas** to streamline data access and integration for **2500+ students across 15+ university chapters**.
- Engineered **REST and GraphQL API endpoints** in TypeScript with **80%+ test coverage** for multiple tables.

Projects

arXiv Semantic Search Engine *May 2026 to Present*
Python, FAISS, BM25, Streamlit, Docker *Demo*

- Built a **hybrid semantic search engine** over **425,000+ arXiv AI/ML/NLP papers**, combining **FAISS dense vector search** with **BM25 sparse retrieval** via **Reciprocal Rank Fusion (RRF)**, achieving **sub-second** corpus retrieval.
- Selected **BAAI/bge-small-en-v1.5** as the sentence encoder after benchmarking against all-MiniLM-L6-v2; generated a **3.5GB corpus** of **384-dimensional dense vectors**.
- Engineered **zero-shot query expansion** via **Llama 3.1 (Groq API)**, prompting the model to generate synonyms, alternate acronyms, and related concepts; layered **LRU caching** to cut end-to-end latency by ∼**50%** on repeated queries.
- Dockerized and deployed on **Hugging Face Spaces** with secrets-managed API tokens; built a **startup-time lazy-loading pipeline** to stream the corpus from HF Hub within a **16GB memory budget**.

Virtual Assistive Robotics Module (Software Engineering Capstone) *January 2026 to May 2026*
Unity, C#, SVN, GitLab *Expo Page*

- Architected a **3D robotic arm simulation** in a **5-person Agile team** for UMD's Neuromotor Control and Learning Laboratory, providing neuroscientists with a platform to study human sense-of-agency (SoA).
- Engineered a **high-throughput data pipeline** to time-synchronize live EEG streams with simulation telemetry at high frequencies (60 to 200+ Hz) to identify neurological markers for SoA. Presented at UMD Engineering Capstone Design Expo 2026.

Skills

Languages: Python, Java, JavaScript, TypeScript, C#, C/C++, SQL
AI/ML: PyTorch, NumPy, pandas, vERL, Modal, scikit-learn, cv2, Matplotlib
Tools: Git/GitHub/GitLab, Docker, HuggingFace Hub, Groq API, Chroma DB, AWS, PostgreSQL