Rohit Gomes

Kolkata, India | +91 9647750262 | gomesrohit92@gmail.com
GitHub | LinkedIn | Portfolio

Summary

AI Engineer driven by a genuine passion for solving real-world problems and a continuous desire to learn and scale within the field. Guided by the core belief that AI should be engineered as a powerful tool to make human work easier and more efficient, rather than a force that replaces human talent. Committed to building practical, impact-driven applications that empower people, streamline everyday workflows, and bring tangible value to real life.

Education

Brainware University, WB
Bachelor of Technology (B.Tech) in Computer Science (AI & ML) *Aug 2022 – Aug 2026*

- Current GPA: 8.7/10.0

St. Stephen's School Dum Dum, WB
Higher Secondary (Class 12, Science Stream) *Completed 2022*

- Academic Performance: 72.25%

St. Stephen's School Habra, WB
Secondary (Class 10) *Completed 2020*

- Academic Performance: 77.80%

Experience

Anti-Aliased CNN-Aided Steering Angle Prediction Autonomous Vehicles
Research Paper Contributor *Sept 2024 – Dec 2024*

- Collaborated on a published research paper developing an anti-aliased Convolutional Neural Network (CNN) architecture to enhance the precision of steering angle prediction in self-driving systems.
- Addressed the "aliasing effect" in conventional CNNs by integrating blur-pooling and low-pass filtering, significantly improving translational invariance and stability under adverse driving conditions.
- Utilized Transfer Learning with VGG-19, ResNet, and DenseNet backbones to optimize feature extraction, achieving superior prediction accuracy on the Sully Chen dataset compared to state-of-the-art models.
- Demonstrated system robustness in challenging scenarios such as low lighting and road obstructions, proving the model's viability for real-world autonomous navigation.

Technical Skills

AI & Systems Architecture: Advanced RAG, Agentic Workflows, Multi-Stage Retrieval, Hybrid Search, Reciprocal Rank Fusion (RRF), Multi-Query Expansion, Few-Shot Learning, Meta-Learning
AI Stack: ChromaDB, Ollama, Google Gemini API, Tiktoken, Cross-Encoders, Meta-Learning Models (MAML)
Languages & Data: Python (Advanced Pandas, cuDF, Modin), SQL, PyTorch, RAPIDS, Parquet
Data Engineering: GPU-Accelerated Pipelines, Semantic Chunking, Metadata Filtering, Vectorization, Dask, Numba
NLP & CV: Text Cleaning, Sentence-Transformers, Anti-Aliased CNN, Object Detection, Classification
Tools & Platforms: Git, Cloudflare, Docker, Streamlit, PyPI Packaging

Projects

Vectra: The Meta-Inference Engine & SDK | *FastAPI, PyTorch, Docker* Aug 2025 – Present

- Architected a full-lifecycle Few-Shot Learning platform that enables training and deployment of high-performance image classifiers using as few as 5 samples per category.
- Solved the "False Positive" problem in standard classifiers by implementing Distance-Based Rejection, using Euclidean distance thresholds to natively detect and reject unknown categories.
- Developed a modular, pip-installable Python SDK to decouple complex inference logic from client applications, facilitating rapid integration into existing production workflows.
- Engineered a real-time cam-inference system with optimized CPU deployment via Docker, ensuring low-latency performance without requiring specialized GPU hardware.

SRE-Pulse: Hybrid Search Engine | *RRF, Vector DB, Python* Sept 2025

- Built a dual-mode retrieval system to bridge the gap between keyword-specific searches and semantic intent, solving the "vocabulary mismatch" problem in traditional RAG.
- Implemented Reciprocal Rank Fusion (RRF) to unify results from BM25 and vector search, ensuring that the most relevant information is prioritized regardless of query structure.
- Optimized system reliability by designing an idempotent data synchronization layer, preventing data duplication while maintaining sub-second retrieval latency.

Scalable Data Engineering & LLM Prep | *RAPIDS, Parquet, NLP* Aug 2025 – Sept 2025

- Engineered high-velocity data pipelines capable of processing millions of unstructured data points, reducing processing overhead by leveraging GPU acceleration.
- Streamlined the transition from "dirty" raw text to model-ready embeddings by building an automated NLP suite that handles complex text normalization and metadata extraction.
- Improved downstream model accuracy by implementing semantic-aware chunking strategies, ensuring critical context is preserved during the vectorization process.

Certifications

- **Getting Started with Artificial Intelligence** – IBM Skills Build (August 2025)
- **Journey to Cloud: Envisioning Your Solution** – IBM Skills Build (August 2025)
- **Introduction to Generative AI** – AWS Educate

Interests

High-Performance Data Engineering, Autonomous Vehicle Systems, Advanced RAG, Agentic Workflows, and Bridging Academic Research with Production Systems.