

SHREERAG NAMBOOTHIRI K H

+91 9188155396 | shreeragkh@gmail.com

Github | LinkedIn

OBJECTIVE

MCA student at Vellore Institute of Technology with hands-on experience in prompt engineering, LLM application development, and AI agent design. Skilled in building Retrieval-Augmented Generation pipelines, multi-step conversational agents, and prompt evaluation workflows using tools such as LangChain, LangGraph, and the Claude API. Seeking an AI/ML internship to contribute to real-world LLM-powered products.

EDUCATION

Vellore Institute of Technology

May 2027 (expected)

M.C.A

Naipunnnya Institute of Management and Information Technology (Univeresity of Calicut)

2022-25

B.sc in Computer Science

PROJECTS

Hybrid Search RAG Application

Python, FastAPI, LangChain, ChromaDB, BM25s, Redis, Streamlit

<https://github.com/shreeragkh/Hybrid-Search-RAG>

- Built a hybrid retrieval-augmented generation pipeline combining BM25s sparse search and ChromaDB dense vector search, merged via Reciprocal Rank Fusion (RRF) and run in parallel using ThreadPoolExecutor.
- Implemented a cross-encoder reranking stage (ms-marco-MiniLM-L-6-v2) on fused candidates to improve retrieval precision before generation.
- Designed a FastAPI backend exposing query, PDF ingestion, authentication, log monitoring, and cache management endpoints, with a Redis caching layer that hashes queries (SHA-256) to skip repeated vector search.
- Developed a Streamlit web UI with separate public and admin views, the latter exposing source attribution, retrieval/timing metrics, raw retrieved chunks, and system health status.
- Integrated Firebase Google Authentication for role-based access control, restricting admin dashboard and management endpoints to an authorized account.
- Added a dynamic document upload flow allowing PDFs to be split, embedded, and indexed into both ChromaDB and BM25 on the fly via the API/UI.
- Containerized the application with Docker and deployed it on an AWS EC2 instance, running the full stack (FastAPI, Streamlit, Redis) end-to-end in a cloud environment

Prompt Tester

Python, LangGraph, Streamlit, Claude API, LLM Evaluation

<https://github.com/shreeragkh/Prompt-Testing>

- Built a prompt engineering benchmarking tool comparing zero-shot, few-shot, and chain-of-thought prompting strategies using the Claude API.
- Used Claude as an LLM-as-judge to score each strategy output on accuracy and clarity (1–10) and automatically determine the winning strategy per run.
- Built session history tracking and CSV export to capture benchmark results across multiple test runs.
- Deployed as an interactive Streamlit app with side-by-side strategy comparison and a win-count bar chart across sessions.

AI to Detect Fake Open Source Projects

- Developed a machine learning model to detect inactive or fake open-source repositories using multi-dimensional GitHub data.
- Engineered features from repository metadata, contributor activity, bot contributions, code similarity, and contribution graphs using Github API.

- Evaluated multiple classification models (including XGBoost and Logistic Regression); Logistic Regression delivered the best performance for the task.
- Applied SHAP explainability to interpret model predictions and identify key influencing features.
- Currently preparing research work for publication.

TECHNICAL SKILLS

Programming	Python, C++
AI/ML	Generative AI, Large Language Models (LLMs), Prompt Engineering, RAG, AI Agents, Semantic Search, Intent Classification, LLM Evaluation, Machine Learning, Pandas, NumPy, Matplotlib
Frontend	React
Prompt Engineering	Zero-shot, Few-shot, Chain-of-Thought, Role-based Prompting, Prompt Optimization, Performance Benchmarking
Frameworks, Libraries	LangChain, LangGraph, FastAPI, Streamlit, Flask, React.js, Scikit-Learn
Retrieval & Search	Hybrid Search (BM25 + Dense), Reciprocal Rank Fusion (RRF), Cross-Encoder Reranking, Semantic Search, Embeddings, Vector Search
Vector Databases & Caching	FAISS, ChromaDB, Vector Embeddings, Redis
Tools, Platforms	Git, GitHub, Firebase Auth
Cloud	AWS EC2, Docker