

Yagyesh Vyas

AI/ML Engineer – LLM Fine-Tuning, RAG and Generative AI

Richmond, VA | (224) 517-8461 | yagyesh612@gmail.com

linkedin.com/in/yagyeshvyas | github.com/yagyeshVyas | yagyesh-portfolio.vercel.app

Education

M.S., Computer Science and IT – University of the Potomac

GPA 3.88/4.0 · Grad Aug 2026

B.E., Computer Science – Gujarat Technological University, India

May 2022

Certifications: IBM AI Applications | IBM NLP | AWS AI and ML Fundamentals | Kubernetes | Generative AI | Information Security

Experience

Data & AI Developer Intern | MKL Management LLC

Feb 2025 – Oct 2025

- Cut 10+ hours of weekly manual work with n8n workflows and Python scripts calling LLM APIs end to end
- Built Power BI and Excel KPI dashboards from spreadsheets, REST APIs, and internal systems; leadership reads them daily
- Automated the processes nobody had time to fix and trimmed recurring operational costs

Projects

VibeGuard: Security and Privacy Scanner for AI-Generated Code | Open source

Node.js, MCP, SARIF

- Free, local-first scanner for AI-built apps: 766 rules across 18 languages and 20+ frameworks, 43 auto-fix rule types, everything on-device
- 13-layer defense: prompt-injection and jailbreak firewall, zero-trust sandbox, runtime interceptors, supply-chain firewall, AES-256-GCM secret vault
- Scope-aware AST taint analysis reaches 86% precision and 96.4% recall on a vulnerability benchmark; results export as SARIF in CI/CD
- 84 MCP tools let Claude Code, Cursor, Windsurf, and other MCP-compatible agents scan and autofix in real time

github.com/yagyeshVyas/VibeGuard

PersonalForge: Open-Source LLM Fine-Tuning Platform

Python, Flask, Unsloth, ChromaDB

- 97 GitHub stars and 11 forks
- No-code fine-tuning: upload documents, pick a Hugging Face model, get a quantized GGUF to run offline
- 17-step data-cleaning pipeline (PII stripping, MinHash dedup, encoding repair, quality gates) handles up to 2M samples from 26 datasets and 6 source types
- rsLoRA, NEFTune, DPO, curriculum learning, gradient checkpointing, and 4-bit quantization squeeze 7B-parameter training onto a free Colab T4
- BGE-M3 embeddings plus ChromaDB RAG with auto-generated chain-of-thought Q&A pairs; supports 58+ models (Llama, Qwen, Mistral, Phi, DeepSeek)

github.com/yagyeshVyas/personalforge

AI Resume Analyzer: ATS Optimization Web App

Python, Streamlit, SQLite

- Live ATS tool: scores resumes against job descriptions for match percentage, missing keywords, and interview odds
- Runs offline via Ollama, LM Studio, or 12+ cloud LLM providers and stores nothing

ai-resume-analyzer-by-yv.streamlit.app

Skills

LLMs and GenAI: fine-tuning (LoRA, rsLoRA, DPO, NEFTune, Unsloth), RAG pipelines, prompt engineering, quantization (GGUF), embeddings (BGE-M3), LangChain, Transformers, Hugging Face, model evaluation

Machine Learning and Programming: Python, JavaScript (Node.js), SQL, Bash, PyTorch, vector search (ChromaDB), data pipelines, MCP tool use, LLM guardrails, OpenAI GPT-4, Claude, Gemini, DeepSeek, Groq

Cloud and DevOps: AWS, Docker, Kubernetes, CI/CD, REST APIs, Git, model deployment, model serving (Ollama), Flask, Streamlit, Playwright, n8n automation, Colab GPU training