Contact

tallhawaheed@gmail.com

www.linkedin.com/in/tallha-waheed-254257115 (LinkedIn)
tallhawaheed.dev/ (Personal)

Top Skills

Data Structures

Big Data

Decision Sciences

Tallha Waheed

AI consultant & Machine Learning Engineer | AI | Data Science | RAG & AI Agents | Generative AI | LLMs | Prompt Engineering | Machine Learning | Python | Building production-grade GenAI systems

Sterling, Virginia, United States

Summary

I am a Machine Learning Engineer and AI consultant (ex-Accenture) helping companies turn LLMs, RAG and AI agents into real products and internal tools.

I work with founders, product teams and tech leaders who want to integrate AI into existing products, automate internal workflows, or validate new AI ideas quickly without building a big in-house team.

I lead a small team of engineers so we can cover end-to-end delivery – from architecture and POCs to production deployments.

What I can help with
• Design & implementation of LLM / RAG systems (retrieval pipelines, vector DBs, evaluation)
• AI agents for workflow automation, tool-calling, integrations with your APIs
• Integrating OpenAI / other LLMs into web apps, backends, CRMs, support tools, etc.
• Architecture reviews, technical roadmaps and AI strategy for your product
• Building prototypes/MVPs for new AI features or internal copilots

How I work
• Project-based consulting & implementation
• Ongoing technical advisory (a few hours per week)
• Independent contractor / staff-augmentation as an ML engineer for AI/LLM teams

If you’re exploring an AI project (or need extra hands on
your ML/LLM work), feel free to DM me here or email:
devtallhawaheed@gmail.com

Experience

Self-employed

Founder & Lead AI Consultant

January 2025 - Present (1 year)

New Jersey, United States

I run a small AI consulting practice focused on LLMs, RAG and AI agents for product teams and businesses.

• Design and implement AI features (chatbots, copilots, search, agents) using LLMs

• Build RAG pipelines with vector databases, retrieval, and evaluation

• Integrate AI into existing products and internal workflows (APIs, backends, SaaS tools)

• Lead a small team of engineers to deliver end-to-end solutions (POC → production)

• Provide technical advisory to founders and teams on AI strategy and architecture

I work with clients as both:

• Consultant / implementation partner for AI initiatives

• Independent ML engineer (contract) embedded into existing teams

Accenture

Sr. Machine Learning Engineer

June 2022 - May 2025 (3 years)

395 9th Ave, New York, NY 10001, United States

- Fine-tuned open-source models (DeepSeek, Mistral, and Llama) using LoRA, PEFT, Adapters, and bfloat16 methodologies, enhancing model personas and aligning brand identity; implemented real-time knowledge updates via RAG on fine-tuned models to ensure accurate and brand-consistent interactions.

- Built automated LLM fine-tuning pipeline, achieving a 23% improvement in GPT-3 responses compared to GPT-4.

- Built interactive agentic workflows using pydantic, Langgraph, llama index and developed backend services in FastAPI, creating RESTful APIs and robust database integrations.

- Developed, deployed, and optimized robust ML models primarily on AWS infrastructure including SageMaker, EC2, and S3.

- Developed and deployed machine learning pipelines leveraging Azure Machine Learning for predictive analytics and automated model deployment.

- Integrated data processing pipelines using Azure Databricks and Synapse Analytics, improving data ingestion speed and accuracy.

- Integrated RAG and Neo4j knowledge graphs, n8n workflows improving financial risk assessment accuracy.

- Architected AI Voice Sytem, reducing conversational latency from 7s to <2s with Twilio WebSockets, custom VAD, and NER models; achieved a 95% cost optimization.

- Developed a credit score model for BMW to optimize decision-making and predict delinquency with 85% accuracy.

- Developed chatbot and recommendation engine for BMW, delivering seamless customer interactions and increasing engagement by 40%.

- Improved LLM response quality by 85% through advanced prompt engineering and RLHF.

- Built intelligent OCR-driven LLM workflows for automated resume screening, reducing shortlisting time by 85%.

- LangChain and LangGraph to orchestrate multiple agents, tools, and functional calling operations, effectively handling complex resume evaluation processes.

- Integrated fine-tuned open-source and API-based autoregressive LLMs (GPT-4 Turbo, LLaMA 2, Falcon-7B) to perform nuanced candidate evaluations.

Dropbox
Machine Learning Engineer
March 2019 - June 2022 (3 years 4 months)
United States

- Credit Scoring Model for a Bank: Developed an AI-driven credit scoring system achieving a 95% accuracy rate in predicting customer defaults. This model incorporated advanced feature engineering, behavioral risk analysis, and model interpretability layers, significantly reducing default rates and enhancing loan approval processes

- Utilized XGBoost, CatBoost, and ensemble methods to build predictive models that classified applicants based on their likelihood of timely repayment.

- Incorporated demographic, behavioral, and transactional data to improve model precision and reduce false positives.

- Delivered comprehensive model documentation, explainability reports, and performance dashboards to ensure transparency and ease of integration with internal systems.

- Engineered predictive time-series models (Prophet, LSTM) and RLHF pipelines, reducing underwriting processes from weeks to minutes.

- Improved document retrieval accuracy by implementing advanced hybrid search methodologies and integrating specialized tools such as Docling, LlamaParser, and Mistral OCR, resulting in faster and more precise document indexing and extraction.

- Engineered an automated content moderation system for a short-form video platform , which reduced negative content by over 85%. Implemented Transformer visions and YOLO models capable of detecting and filtering inappropriate content.

- Built a short video recommendation model increasing user engagement time by up to 50%. Leveraged collaborative filtering and content-based filtering techniques to suggest relevant videos to users.

Incedo Inc.

Data Scientist

May 2016 - February 2019 (2 years 10 months)

Santa Clara County, California, United States

· Developed an LLM-enhanced AI search engine with knowledge graph-powered hierarchical ranking, semantic search, and auto-suggestions, significantly improving search recall and precision.

· Integrated retrieval-augmented generation (RAG) pipelines with LLMs for contextualized and query-specific search responses.

· Designed high-availability search APIs with Redis caching and ElasticSearch, enabling low-latency, real-time search.

· Optimized the ranking algorithm by implementing TF-IDF and transformer-based embedding (BERT, RoBERTa), leading to 30% better relevance scoring for search results.

· Automated indexing and retrieval processes, reducing data ingestion time by 50% and improving query performance.

· Implemented an unsupervised topic modeling pipeline using Latent Dirichlet Allocation (LDA), Singular Value Decomposition (SVD), and Probabilistic Models, categorizing news articles into predefined topics (e.g., Sports, Politics, and Finance).

· Optimized distributed inference pipelines on Azure DataBricks, enabling large-scale processing of real-time news data.

· Enhanced clustering accuracy by incorporating probabilistic distribution matrices, improving article classification precision by 32%.

· Integrated topic modeling outputs into downstream recommendation systems, refining personalized content delivery.

· Collaborated with a team of data scientists to design and develop an intelligent medical chatbot for real-time patient assistance, including appointment scheduling, automated medical record management, and symptom-based guidance.

· Engineered a Transformer-based NLP pipeline, integrating BI-Encoder and Cross-Encoder architectures to enhance conversational accuracy and response relevance.

· Integrated OCR-based document processing, enabling automated extraction of prescriptions, medical records, and diagnostic reports.

Education

Becker College

Bachelor of Science - BS, Computer Science · (June 2013 - July 2017)