MANSOOR HAYAT, MBA | PhD | P.Eng.

Agentic AI Engineer | Prompt Engineering | LLM Orchestration | Workflow Automation

+1-204-390-5379 | Mansoor.Hayat@umanitoba.ca | linkedin.com/in/engr-mansoor-hayat | github.com/MansoorHayat777 |
Open to Relocate: Montreal, Vancouver, Toronto, San Francisco

Current Immigration Status: Closed Work Permit (Employer-Specific) | Valid: Sep 27, 2027 | Employer: University of Manitoba

PROFESSIONAL SUMMARY

Production-focused Agentic AI Engineer and NSERC-funded researcher with PhD in Data Science and 7+ years of experience designing, building, and shipping AI systems at scale. Deep hands-on expertise in RAG pipeline design, LLM API integration (GPT-4, LLaMA-3, Claude/Anthropic), vector database embedding retrieval (FAISS, Pinecone), LangChain/LlamaIndex orchestration, advanced prompt engineering, and agentic multi-step workflow automation deployed in containerised production environments. Track record: 70% reduction in retrieval time, 35% reduction in analyst workload, 3x throughput improvement, sub-100ms inference latency at enterprise scale. 27 peer-reviewed publications.

TECHNICAL SKILLS

RAG & LLM Systems: RAG pipeline design and optimisation; LangChain, LlamaIndex, FAISS orchestration; embedding-based retrieval; vector databases (FAISS, Pinecone, Chroma); GPT-4, LLaMA-3, Mistral, Claude/Anthropic APIs

Advanced Prompt Engineering: Zero-shot, few-shot, chain-of-thought, tree-of-thought, self-consistency, system prompt design, persona engineering, structured output prompting, prompt chaining, context window management

Agentic AI & Orchestration: Multi-step agentic workflows; LangChain agents; tool use and function calling; multi-agent orchestration; n8n workflow automation (no-code and hybrid pipelines); automated enterprise workflow integration; ReAct, Plan-and-Execute patterns

LLM Governance & Safety: LLM governance controls, prompt logging, cost tracking per inference call, latency monitoring, hallucination detection pipelines, responsible AI operation within PHIA and PIPEDA regulatory requirements

LLM Engineering: Fine-tuning (LoRA, QLoRA, SFT, instruction tuning); RLHF concepts; model quantisation (GPTQ, AWQ); ONNX export; automated evaluation benchmarks; drift detection and rollback strategies

Cloud & MLOps: GCP (Vertex AI, Cloud Run, BigQuery), Azure ML, AWS SageMaker; Docker, Kubernetes, GitHub Actions CI/CD; MLflow, DVC, Apache Airflow, Kafka

Languages & Frameworks: Python (primary), SQL, Bash; PyTorch, Hugging Face, Scikit-learn, OpenCV

PROFESSIONAL EXPERIENCE

Agentic AI Engineer & Applied AI Researcher | Postdoctoral Research Fellow

University of Manitoba, Section of Neurosurgery | Winnipeg, MB **Sep 2024 - Present**

- Designed and shipped a production RAG-based clinical data extraction system combining LLM APIs (GPT-4, LLaMA-3), FAISS vector retrieval, and LangChain orchestration - achieving 3x throughput over manual baseline. Deployed as containerised REST API on GCP Cloud Run with 24/7 uptime.
- Implemented multi-step agentic workflows for automated clinical data processing using n8n for workflow orchestration and LangChain for LLM agent chaining: agents performing document parsing, entity extraction, validation, and structured output generation, reducing end-to-end processing time by 60%.
- Engineered advanced prompt engineering strategies (chain-of-thought, structured output, few-shot with clinical examples, context window management) to maximise extraction accuracy on medical device data across 150,000 annotated records.
- Built LLM fine-tuning pipelines (LoRA, QLoRA) on domain-specific clinical corpora with automated evaluation benchmarks validating extraction accuracy against annotated ground truth, enabling safe versioned deployment with automated rollback on GCP Vertex AI.
- Implemented LLM governance controls: prompt logging, cost tracking per inference call, latency monitoring, and hallucination detection pipelines to ensure responsible, auditable AI operation within PHIA and PIPEDA regulatory requirements.

- Secured NSERC Discovery Grant (I2IPJ 598125-24, $35K-$50K/yr) - demonstrating end-to-end ownership from research through production deployment.

Senior ML Engineer & AI Consultant - LLM & Agentic Systems

Softbox Technologies | Remote **Aug 2022 - Sep 2024**

- Designed and shipped enterprise RAG agents (LangChain, FAISS, GPT-4/LLaMA-3) for financial document Q&A and intelligent enterprise search, achieving 70% faster retrieval and 35% reduction in analyst workload. Deployed as containerised microservices on GCP Cloud Run with CI/CD via GitHub Actions.
- Built LLM fine-tuning and evaluation platforms (LoRA, QLoRA) on domain-specific enterprise corpora with model registry, automated evaluation, drift detection, and rollback strategies on GCP Vertex AI.
- Developed multi-agent orchestration pipelines for enterprise workflow automation: tool-use agents integrating with enterprise APIs, document management systems, and databases; reducing manual processing by 40%.
- Applied advanced prompt engineering (zero-shot, few-shot, chain-of-thought, system prompt design) across GPT-4, LLaMA-3, and Mistral deployments; built structured output validation pipelines to ensure reliable JSON/schema-compliant LLM outputs.

Doctoral Researcher - AI & Computer Vision

Chulalongkorn University | Bangkok, Thailand **Aug 2021 - Jul 2024**

- Implemented explainable AI modules (Grad-CAM, SHAP) and uncertainty estimation directly applicable to responsible AI governance. Published 6 first-author papers in IEEE Access and IEEE EMBC. Best Paper Award IEEE TENCON 2023.

Lecturer & Lab Engineer

University of Sargodha | Pakistan **Jan 2015 - Apr 2021**

- Built NLP automation pipelines for text classification and information extraction; supervised 20+ engineering project teams.

AI/ML Technical Writer (Freelance)

BuffML | Remote **2021 - Present**

- Publishes practitioner articles on LLMs, RAG, MLOps, and prompt engineering - translating CVPR, NeurIPS, and ICLR research into applied engineering content.

SELECTED AGENTIC AI & RAG PROJECTS

- Enterprise RAG Agent (LangChain, FAISS, GPT-4): 70% faster retrieval; 35% analyst workload reduction; containerised microservice on GCP Cloud Run; CI/CD pipeline; structured output validation.
- LLM Fine-tuning Platform (LoRA/QLoRA, 2022-24): Domain-specific fine-tuning; automated eval and rollback; GCP Vertex AI; model registry and versioning; cost and latency monitoring.
- Clinical Data RAG Pipeline (NSERC-funded, 2024-25): 150,000 annotated records; FAISS vector retrieval; LangChain orchestration; chain-of-thought prompt engineering; LLM governance; 3x throughput; 24/7 production.
- Agentic Workflow Automation (UofM, 2024-25): Multi-step document parsing and extraction agents (n8n + LangChain); 60% manual workload reduction; PHIA-compliant; containerised REST API; audit logging.

EDUCATION

- Doctor of Philosophy, Data Science / ML / AI (Distinction) | Chulalongkorn University, Bangkok | 2021-2024
- Master of Business Administration, Accounting & Finance | NCBA&E, Lahore, Pakistan | 2018-2022
- Master of Science, Electrical Engineering | ISP Multan, Pakistan | 2016-2018
- Bachelor of Science, Electrical Engineering | UET Taxila, Pakistan | 2011-2015

AWARDS, CERTIFICATIONS & PROFESSIONAL STANDING

- NSERC Discovery Grant (I2IPJ 598125-24, 2025-2026) | Research Manitoba Postdoctoral Fellowship (2025-2026)
- Best Paper Award, IEEE TENCON 2023 | PhD with Distinction (2024) | 27 peer-reviewed publications
- ML Specialisation (Stanford/DeepLearning.AI) | MLOps Specialisation (DeepLearning.AI) | Mathematics for ML: Linear Algebra, Multivariate Calculus, PCA (Imperial College London)
- AI for Medicine Specialisation (DeepLearning.AI) | Deep Learning for Healthcare (University of Illinois) | Explainable Deep Learning for Healthcare (University of Glasgow)

- AWS ML & Computer Vision | Big Data Specialisation (UC San Diego) | Azure AI Applications
- P.Eng. - Engineers Geoscientists Manitoba (APEGM) | Registered Engineer, Pakistan Engineering Council
- PHIA Certificate MB74615 (University of Manitoba) | TCPS 2: CORE 2022 (Research Ethics) | IOSH Managing Safely | WHMIS

REFERENCE

- Dr. Nouman Ahmad | Researcher / AI & Data Science | NAISS Sweden | nouman.ahmad@naiss.se