Shivam Mishra

AI Engineer

✉ shivammishrrr@gmail.com ☎ 9054540509 Bengaluru, India shivammishra27 shivammishrr

Profile

AI Engineer with over 1.5 years of experience deploying generative AI systems in production environments across multiple industries. Specialized in RAG systems, document parsing, multi-agent architectures, and LLM fine-tuning. Proven track record of implementing solutions for the Saudi Ministry of Education, Sweden's largest fishing e-commerce platform, and healthcare information systems. An infinite learner, constantly working on diverse projects across various domains and solving real-world problems through applied AI.

Professional Experience

AI Engineer, *Neuramonks* 06/2024 – present

RFP Evaluation System (Saudi Ministry of Education)

- **Addressed** complex proposal evaluation needs by implementing a **LangGraph multi-agent** system that successfully processed documents exceeding standard context windows
- **Engineered** robust **document parsing** pipeline with fallback **OCR**, standardizing **PDF/DOCX** files into structured Markdown format for consistent LLM processing
- **Transitioned** from OpenAI PoC to production by deploying **Qwen** models with **LM-Enforcer**, resulting in structured **JSON** outputs that satisfied strict government privacy requirements

Fiske: Intelligent Fishing Assistant (The Largest Ecommerce for Fishing Equipment in Sweden)

- **Built** intelligent chatbot with multi-agent architecture that delivered personalized product recommendations by leveraging Cohere's **multilingual reranker** (rerank-multilingual-v3.0) to optimize retrieval results
- **Solved** product discovery challenges by creating an advanced RAG system with Pinecone and **text-embedding-3-large**, achieving **80% improvement** in retrieval accuracy across **30K+ fishing products**
- **Engineered** bilingual semantic parsing engine that accurately processed **English/Swedish** queries with **text-to-SQL** capabilities, driving increased customer engagement

AI Engineer, *GoAskNow* 12/2023 – 05/2024

Multimodal Image Processing Pipeline

- **Resolved** image processing limitations by engineering a production-ready image-to-text system on GCP with **Gemini** API integration, enhancing client workflow capabilities

HR Policy Assistant Chatbot

- **Improved** HR operations efficiency by designing a **RAG**-based chatbot that reduced response times for employee inquiries through accurate document parsing
- **Deployed** scalable AI solutions using Flask, **GCP**, **Docker**, and **Amazon Lambda** with proper monitoring for production environments

AI Engineer-Intern, *Mokx* 10/2023 – 12/2023

"Arya" - the world's first AI Acharya

- **Solved** knowledge accessibility challenges by architecting "Arya" on **Azure** with a **specialized RAG** system that improved information quality through advanced parsing and embedding of Hindu scriptures (**Vedas, Shlokas**), resulting in **52%** better response relevance.

Skills

Programming, Infrastructure & MLOps

- Python · R · SQL
- LangChain · LangGraph · CrewAI · Haystack
- AWS · GCP · Azure
- Docker · Kubernetes · CI/CD · MLflow · Model monitoring
- Transformers · Datasets · Spaces

Generative AI & Retrieval Systems

- LoRA · QLoRA · RLHF pipelines
- Chunk-aware retrieval · Hybrid search · HyDE · Reranking
- Pinecone · Qdrant · Neo4j
- Text-to-image · Diffusion models

ML Architectures & Optimization

- Transformers (encoder-decoder, causal)
- Mixture-of-Experts (MoE)
- Rotary Positional Embeddings (RoPE)
- PPO · DPO · DRPO
- GPTQ · AWQ · FlashAttention-2 · vLLM · Unsloth

Applied NLP & CV Systems

- Summarization · Text-to-SQL · Intent detection · Structured output
- Task routing · Context-aware agents
- OCR · Image-to-text · Vision-language models
- Retrieval metrics · Model benchmarking · LangSmith · Trulens · OpenAI Eval

Projects

DocTrAI, *Deep-Research based multi-tool assistant* ↗

- **Core Features**: Instant AI-powered health insights (**symptoms, medications, lifestyle**) using Groq's fast models.
- **Advanced Research Capabilities**: **Deep-Research** Mode with collaborating AI agents and a Smart Research System that accesses sources like **Wikipedia**, **Arxiv**, **Google Search**, and **Firecrawl** to deliver evidence-based reports in minutes instead of hours.

Sarvagya, *Manus AI open source version*

- **Developed** an open-source autonomous AI agent inspired by **Manus AI**, utilizing a modular multi-agent architecture to execute complex tasks like data analysis and content generation.
- **Integrated** tool invocation (e.g., web browsing, code execution), facilitating seamless interaction with external applications.

Finetuned llama 3 and Incorporated in RAG

- **Fine-tuned Llama 3** model on custom dataset using **PEFT LoRA** technique for HR-related question answering.
- **Implemented** a **RAG-based** system with vector storage, **Llama Index** for information retrieval, and **LangChain** for language processing tasks.

Education

Bachelors of Engineering, *Vishwakarma Government Engineering College* 2019 – 2023

Certificates

- Deep Learning Specialization ↗
- Machine Learning Specialization ↗
- Google Data Analytics ↗