

PRIYANSHU YADAV

AI Engineer · Large Language Models (LLMs) · Deep Learning · RAG Systems · MLOps

+91-8447903651 | priyanshu.py98@gmail.com | linkedin.com/in/priyanshuyadav98 | github.com/PriyanshuGeTRekT | New Delhi, India

Professional Summary

AI Engineer with 3+ years of hands-on experience in artificial intelligence, machine learning, and deep learning engineering, specialising in Large Language Models (LLMs), Natural Language Processing (NLP), and Retrieval-Augmented Generation (RAG) systems. Currently designing and deploying LLM-powered applications at RazorInfoTech using GPT-4, Claude, and LLaMA, including an enterprise HRMS chatbot with tool calling and RAG. Deep expertise in prompt engineering, fine-tuning, and transformer architectures such as BERT, the GPT series, and LLaMA. Proficient in Python, PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LangGraph, and LlamaIndex. Experienced in building NLP solutions for text generation, summarization, conversational AI, and RAG pipelines, and in deploying and self-hosting models in production with MLOps best practices: CI/CD pipelines, LLM observability (Langfuse), model monitoring, and version control. Hands-on experience with cloud AI services across AWS, Microsoft Azure, and GCP. Strong track record of cross-functional collaboration with data scientists, software engineers, and product managers to deliver scalable AI/ML solutions for real-world client problems.

Core Competencies

Large Language Models (LLMs) | Retrieval-Augmented Generation (RAG) | Natural Language Processing (NLP) | Deep Learning
Transformer Architectures (BERT, GPT, LLaMA, T5) | Prompt Engineering | Fine-Tuning | Transfer Learning
LLM Tool Calling (Function Calling) | AI Chatbot Integration | LLM Observability | Self-Hosted LLM Serving (vLLM)
LangChain | LangGraph | LlamaIndex | Hugging Face Transformers | OpenAI APIs | Anthropic API
MLOps | CI/CD Pipelines | Model Monitoring | Version Control | Model Deployment | Model Optimization
AWS (Amazon Web Services) | Microsoft Azure | GCP (Google Cloud Platform) | Docker | Kubernetes
Text Generation | Summarization | Conversational AI | Semantic Search | Vector Databases (FAISS, Qdrant, Pinecone)
Hyperparameter Tuning | Model Scaling | Low-Latency Inference | Agentic AI Workflows | AI Automation

Technical Skills

LLM & Deep Learning: Large Language Models (LLMs), transformer architectures (BERT, GPT-3.5, GPT-4, LLaMA, T5), deep learning, fine-tuning, transfer learning, hyperparameter tuning, prompt engineering (few-shot, zero-shot, chain-of-thought, system prompt design)
AI Frameworks & Libraries: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LangGraph, LlamaIndex, scikit-learn, NumPy, Pandas, OpenAI APIs, Anthropic API, Langfuse, vLLM
NLP & Generative AI: Natural Language Processing (NLP), Retrieval-Augmented Generation (RAG), text generation, summarization, conversational AI, AI chatbots with tool calling (function calling), named entity recognition, semantic search, multi-turn dialogue systems
RAG Pipeline: Source extraction, sanitisation, chunking strategies, embedding generation, vector indexing, semantic retrieval, re-ranking, hallucination detection, citation accuracy validation
Vector Databases: FAISS, Qdrant, Pinecone, Weaviate, pgvector, hybrid search, embedding models
LLM Serving & Observability: Self-hosted LLM deployment on cloud GPUs (vLLM, Hugging Face TGI), Langfuse observability (request tracing, token & cost tracking, latency/quality monitoring), evaluation pipelines, guardrails, tool/function-calling API design
MLOps & Deployment: MLOps, CI/CD pipelines (GitHub Actions), model monitoring, version control (Git), Docker, Kubernetes, production model deployment, model optimization, low-latency inference, model scaling
Cloud AI: AWS (EC2, S3, SageMaker, IAM), Microsoft Azure (Azure OpenAI, Azure ML, Azure Entra ID), GCP (Google Cloud Platform), cloud AI infrastructure management
Programming: Python (3+ years), JavaScript (ES6+), TypeScript, SQL, Bash/Shell scripting
Backend & APIs: FastAPI, Django, Node.js, Express.js, REST APIs, GraphQL, OAuth 2.0, JWT, async API design
Frontend: React.js, Next.js, TypeScript, Redux, HTML5, CSS3, webpack
Databases: MySQL, PostgreSQL, MongoDB, SQLite, BigQuery
Tools & Practices: Git, GitHub, JIRA, Postman, Agile/Scrum, technical documentation, cross-functional collaboration

Work Experience

AI Architect | RazorInfoTech

Jun 2026 – Present · New Delhi, India

AI Solutions & Automation: LLM Application Design, Tool-Calling Agents, RAG, Intelligent Workflows

- Integrated an enterprise AI chatbot into the company's HRMS software using RAG and LLM tool calling. Built retrieval pipelines over HR policies for grounded, citation-backed answers and designed function-calling interfaces

that execute live HRMS actions (leave balances, attendance, ticket creation) with role-based access controls and structured output validation.

- Design and deploy AI solutions built on Large Language Models (LLMs) such as GPT-4, Claude, and LLaMA. Architect multi-turn conversational AI systems and text generation pipelines on transformer architectures, applying prompt engineering strategies (chain-of-thought, few-shot, role prompting) tuned to each domain and use case.
- Host and operate open-source LLMs on cloud GPU infrastructure with optimised inference serving (vLLM) for low-latency, high-volume traffic. Implemented and maintain end-to-end LLM observability (Langfuse tracing, token and cost tracking, latency and answer-quality monitoring, evaluation loops) alongside MLOps practices such as CI/CD, version control, and model monitoring.
- Build and fine-tune LLM applications using LangChain, LangGraph, and OpenAI APIs. Integrate Hugging Face Transformers for model selection, embedding generation, and domain-specific NLP tasks, and apply transfer learning and hyperparameter tuning to adapt pre-trained models for client use cases.
- Deliver AI-powered automations for cross-functional teams (HR, operations, sales, support). Partner with data scientists, engineers, and product managers to identify high-impact workflows, then ship automations for document processing, reporting, and repetitive back-office tasks that cut manual effort and turnaround time.

AI Full Stack Developer (Intern) | Webisdom Management Services Pvt. Ltd.

Jan 2026 – May 2026 · New Delhi, India

Digital Transformation Consulting: SaaS Products, AI-Integrated Platforms, CRM Systems

- Designed and implemented end-to-end Retrieval-Augmented Generation (RAG) pipelines in Python using LangChain and LlamaIndex, covering the full AI/ML workflow: source content extraction, sanitisation, semantic chunking, embedding generation via Hugging Face sentence-transformers and OpenAI APIs, FAISS vector indexing, semantic retrieval with re-ranking, and structured prompt engineering with robust error handling.
- Built agentic LLM workflows with LangChain/LangGraph on transformer architectures. Designed system prompts, few-shot templates, and chain-of-thought reasoning structures; evaluated model outputs for hallucination and citation accuracy; applied model optimisation to reduce latency and improve response quality; monitored pipelines with Langfuse (LLM tracing, token usage, retrieval quality).
- Developed conversational AI features and NLP-powered text generation and summarization using OpenAI APIs and Anthropic API, integrating them into production backend systems (Python, FastAPI) with async request handling, rate limit management, and structured output parsing.
- Implemented MLOps practices for AI pipeline management: CI/CD pipelines (GitHub Actions), Docker containerisation, version control (Git), and model monitoring on AWS production infrastructure. Executed zero-downtime database migrations with Python/SQL automation scripts.
- Architected full-stack platforms (React/Next.js, TypeScript, Node.js, MySQL/MongoDB); implemented OAuth 2.0, JWT, and RBAC; led intern team of 4; achieved 6/6 KRAs in February 2026 (5 rated Excellent), 100% on-time delivery rate.

Web Developer | Ascent Edge LLP

May 2024 – May 2025 · Remote

Data-Driven Performance Marketing Agency

- Built the MLOps pipeline for the agency's attribution and conversion-prediction models: Dockerised training and inference, GitHub Actions CI/CD for automated retraining and deployment, and Git-versioned model artefacts on AWS.
- Instrumented deployed models with monitoring for drift, latency, and prediction quality, with alerting on degradation, and tuned inference for low-latency scoring in live campaigns.
- Developed tracking APIs and attribution data pipelines feeding those models; ran Python (Pandas) and SQL analysis and reported through Looker Studio dashboards.
- Built high-performance web applications (React, TypeScript, HTML5, CSS3) achieving 35% faster page loads and an 18% conversion uplift via A/B testing and Webpack optimisation.

Software Developer | eSentire

May 2023 – Apr 2024 · Toronto, Canada
(Remote)

Global Cybersecurity Leader in Managed Detection & Response (MDR) | XDR Platform Engineering

- Built Python-based NLP and deep learning classification pipelines for threat alert processing, applying transformer-based text classification models, Pandas/NumPy data preprocessing, and scikit-learn evaluation frameworks to automate alert routing and reduce manual processing by ~20%.
- Implemented MLOps practices for internal AI tooling: CI/CD pipelines (GitHub Actions), version control (Git), model monitoring dashboards, and Docker-based deployment on AWS cloud infrastructure (EC2, S3, IAM), ensuring production model reliability and low-latency performance.
- Collaborated cross-functionally with data scientists, software engineers, and stakeholders to design and deliver AI/ML solutions. Developed and maintained backend REST APIs (Python, Node.js), wrote unit and integration tests (TDD), and conducted code reviews aligned with engineering best practices.

- Administered Microsoft Azure-based infrastructure (Azure Entra ID, Azure AD, IAM permissions); applied data analysis with Python (Pandas, NumPy, scikit-learn) and SQL to operational datasets; produced structured stakeholder reports and BI dashboards for senior leadership.

Software Developer | [SPF OPC Pvt. Ltd.](#)

Aug 2022 – Mar 2023 · Bhopal, India

- Developed and maintained web applications; executed QA and security testing protocols; collaborated on system architecture design with version control (Git) and CI/CD workflows.

Business Analyst Intern | [SSF Seneca College](#)

Dec 2021 – Mar 2022 · Toronto, Canada

- Produced structured data analysis reports and stakeholder presentations; increased operational efficiency by 20% and stakeholder consensus by 30% through data-driven process improvement.

Projects

End-to-End RAG Pipeline with Transformer

Jan 2026 – May 2026

Models | [NamanDarshan.com](#)

- Designed and shipped a production Retrieval-Augmented Generation (RAG) pipeline in Python (LangChain, LlamaIndex, Hugging Face Transformers, PyTorch, OpenAI APIs). The workflow covers source content extraction, sanitisation, semantic chunking, embedding generation with transformer-based sentence models, FAISS vector indexing, semantic retrieval with re-ranking, prompt engineering, and structured output generation. Applied model optimisation for low-latency inference; monitored with Langfuse for hallucination detection and citation accuracy.
- Deployed on AWS with Docker containerisation, a GitHub Actions CI/CD pipeline, and MLOps monitoring. Full-stack platform with React/Next.js TypeScript frontend, FastAPI Python backend, MySQL database, and OAuth 2.0 authentication.

LLM-Powered NLP Classification & Agentic Pipeline

May 2023 – Oct 2023

- Built a Python agentic NLP workflow (LangChain, Hugging Face Transformers, scikit-learn) for threat alert text classification and routing. Applied transformer-based deep learning models, few-shot prompt engineering, chain-of-thought reasoning, and structured output parsing; evaluated model outputs for accuracy and ran MLOps practices (CI/CD, version control, model monitoring) on Microsoft Azure infrastructure.

Conversational AI CRM Platform

Jan 2026 – May 2026

- Engineered a full CRM (150+ daily users) with a Python/FastAPI backend, conversational AI notification system, and NLP-powered data processing pipelines. Used PyTorch and Pandas/NumPy for data transformation; maintained GitHub Actions CI/CD and Docker deployment with version control best practices.

AI Question Paper Generator API

Jul 2025

- Python REST API (FastAPI, SQLite) that generates structured question papers using NLP topic modelling, difficulty classification, and distribution weighting. Packaged with Docker; demonstrates Python AI engineering and model serving principles applicable to LLM application development.

Education

Master of Science (M.Sc.) in Computer Science, Liverpool John Moores University

Feb 2023 – Dec 2025

Post Graduate Diploma in Software Development, IIIT Bangalore

Feb 2023 – Mar 2024

Post Graduate Diploma in International Business Management, Seneca College, Toronto, Canada

Sep 2021 – May 2022

Bachelor of Science (B.Sc.) in Computer Science and Hotel Administration, IHM PUSA, New Delhi

Aug 2016 – May 2019

Certifications and Achievements

- AI Architect, RazorInfoTech (Jun 2026):** Designing LLM-powered applications with GPT-4, LLaMA, and LangChain; HRMS AI chatbot integration with tool calling and RAG, self-hosted LLM serving with observability (vLLM, Langfuse), and cross-team AI automation for enterprise clients.
- #1 in feature delivery at Webisdom (Jan–May 2026):** 6/6 KRAs achieved (5 rated Excellent), 100% on-time delivery, same-day critical bug resolution.
- Production RAG pipeline shipped** (LangChain, Hugging Face, FAISS, Langfuse): full AI/ML workflow from source extraction through transformer-based embedding to hallucination-evaluated output.
- LLM-powered NLP classification pipeline at eSentire:** transformer text classification, few-shot prompting, structured output parsing, MLOps deployment on Azure cloud infrastructure.
- 18% conversion uplift and 35% page speed improvement at Ascent Edge:** A/B testing, Python analytics, and CI/CD deployment pipelines.