Diwita Banerjee

AI/ML Engineer

+17035856489 | diwitabanerjee01@gmail.com | LinkedIn | GitHub | Open to Relocate

PROFESSIONAL SUMMARY

AI/ML Engineer with 5+ years of experience designing and deploying production-grade machine learning solutions in financial services, healthcare, and enterprise domains. Skilled in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), and MLOps. Experienced in fine-tuning Transformer models, building scalable APIs, and deploying AI systems on AWS, GCP, and Azure. Delivered LLM-powered pipelines that improved document retrieval accuracy by 30% and reduced inference latency by 40%. Automated ML workflows with Airflow, MLflow, Jenkins, and Docker, cutting deployment cycles by 50%. Strong background in regulatory-compliant AI, feature engineering, and real-time analytics. Adept at collaborating with cross-functional teams to drive business impact through AI adoption.

TECHNICAL SKILLS

- **Languages & IDEs:** Python, SQL, Shell Scripting, MATLAB, Jupyter Notebook, VS Code, Google Colab, SSMS
- **Frameworks & Libraries:** Pandas, NumPy, Scikit-learn, TensorFlow, Keras, PyTorch, XGBoost, Matplotlib, Seaborn, NLTK, spaCy, Sentence-BERT, Hugging Face Transformers, LangChain
- **Machine Learning:** Logistic Regression, Linear Regression, Decision Trees, Random Forests, SVM, Naive Bayes, A/B Testing, Model Evaluation, Feature Engineering
- **Deep Learning & NLP/LLMs:** CNN, RNN, LSTM, BERT, RoBERTa, GPT-3/4, LLMs, Prompt Engineering, RAG pipelines, Fine-tuning (LoRA/PEFT)
- **MLOps & Deployment:** Docker, Kubernetes (EKS), FastAPI, Flask, MLflow, Jenkins, Terraform, Airflow, PySpark, Model Drift Detection, Usage Metrics Dashboards
- **Cloud & Visualisation:** AWS (Lambda, EC2, S3, RDS, SageMaker, SQS, SNS, CodeDeploy, CloudWatch, API Gateway), GCP, Azure, Tableau, Power BI
- **Databases & Tools:** PostgreSQL, MySQL, MongoDB, Cassandra, Redis, Neo4j, SQL Server, SQLAlchemy
- **Statistical Techniques:** Hypothesis Testing, Data Modeling, Data Visualization, Experiment Design, A/B Testing
- **Collaboration Tools:** Agile Development, Requirements Gathering, Stakeholder Engagement, AI Insights Visualization

PROFESSIONAL EXPERIENCE

Freddie Mac | AI/ML Engineer **Jun 2024 – Present**

- Designed and implemented LLM-powered RAG pipelines using FAISS and Hugging Face, increasing financial document retrieval accuracy by 30%.
- Built and deployed FastAPI-based microservices on AWS EKS and Lambda, reducing model inference response time by 40% and improving scalability.
- Fine-tuned Transformer models, including BERT and RoBERTa, on compliance-related datasets, improving contextual relevance by 25%.
- Automated end-to-end ML lifecycle management using Airflow, MLflow, Jenkins, and Docker, shortening deployment cycles by 50%.
- Developed monitoring dashboards with PySpark and AWS CloudWatch for drift detection and performance analysis of deployed models.
- Collaborated with quantitative analysts and compliance officers to deliver regulator-compliant AI solutions with transparent auditability.

George Mason University | AI Research Assistant **March 2024 – Nov 2024**

- Built fairness and bias evaluation pipelines for multimodal AI models, including Stable Diffusion, LLaVA, and InstructBLIP, uncovering demographic disparities.
- Developed a medical VQA system using MIMIC-IV EHR data that reduced clinician query resolution time by 25% in controlled trials.
- Optimised large-scale training and inference workflows on the Hopper HPC cluster, reducing runtime by 40%.
- Conducted federated multimodal experiments for privacy-preserving healthcare AI and improved model generalization in clinical settings.
- Benchmarked multiple Vision-Language Models (VLMs) and contributed results to academic publications.
- Co-authored ongoing research papers in responsible and federated healthcare AI.

Cognizant Technology Solutions | Machine Learning Developer, Salesforce CRM **Nov 2021 – Jul 2023**

- Built predictive lead/case/renewal scoring models with Einstein Prediction Builder and automated nightly training and scoring via Apex batch and queueable jobs.
- Delivered prescriptive recommendations using Einstein Discovery and Next Best Action, boosting SLA compliance and agent productivity.
- Develop CRM Analytics datasets and dashboards to monitor win-rate lift, adoption, precision/recall, and model drift, with alerts for threshold deviations.
- Integrated external ML endpoints through MuleSoft and Named Credentials with retry, timeout, and audit logging, ensuring secure and reliable data exchange.
- Implemented event-driven and batch pipelines with Platform Events and REST APIs for near real-time updates and high-volume processing.
- Automated CI/CD pipelines with Git and Copado, added unit/integration tests for high coverage, and authored runbooks to streamline monitoring and rollback.
- Recognized with the “Working as One” (2023) and “Always Striving & Never Settling” (2022) awards.

Larsen & Toubro Infotech | AI Developer – Full Stack Applications **Jun 2020 – Oct 2021**

- Engineered AI-driven predictive analytics modules using Spring Boot, Python, and Oracle SQL, reducing backend response latency by 39%.
- Developed intelligent data-driven UI components with React.js, improving usability and long-term scalability.
- Automated defect root cause analysis workflows, reducing issue resolution time by 30%.
- Designed APIs for scalable integration of AI models into enterprise applications.
- Collaborated with business stakeholders to align AI feature development with operational requirements.
- Integrated backend analytics with enterprise dashboards for real-time monitoring and decision-making.

RESEARCH & PROJECTS

MedBLIP – Medical Image Captioning (GMU, Spring 2025)

- Fine-tuned BLIP on the ROCO dataset for medical imaging captioning tasks.
- Achieved benchmark improvements across CIDEr, SPICE, and BERTScore.
- Compared performance with BLIP-2, InstructBLIP, and Gemini models.

CloudMart—Multicloud AI E-Commerce Platform (Bootcamp, Mar 2025)

- Designed and deployed a GPT-powered shopping assistant across AWS, GCP, and Azure using Terraform, Kubernetes, and Docker.
- Implemented real-time product discovery, intelligent Q&A, and analytics pipelines.
- Built cross-cloud ETL-style AI pipelines with Git-based CI/CD, improving scalability and observability.

News Summarisation & Simplification (GMU NLP Project, Fall 2024)

- Built an LSTM-based summarizer and T5 simplification model using Hugging Face Transformers and TensorFlow.
- Trained on the WikiAuto dataset and achieved BLEU 0.22 and Flesch-Kincaid readability 8.31.
- Delivered an end-to-end NLP pipeline to improve accessibility and comprehension of complex news content.

Movie Recommendation (IJARESM, 2022)

- Developed a BERT-based semantic search and recommendation engine on the CMU Movie Summaries dataset.
- Enhanced semantic retrieval and recommendations with a 15% improvement in accuracy.
- Published results in the IJARESM research journal.

EDUCATION

Master of Science in Computer Science (Machine Learning Concentration) **Aug 2023 – May 2025**
George Mason University, Fairfax, VA

- Coursework: Algorithms, AI, Data Mining, Machine Learning, Deep Learning, Advanced NLP
- Honours & Activities: Led and Won 2nd Position in GDG-GMU Hackathon; research on RCRS for Prompt Injection Mitigation