SUMMARY

- **AI Engineer** with **4 years of experience** developing and deploying ML systems across **finance, cybersecurity,** and **enterprise AI.**
- Orchestrated AI pipelines integrating conversational AI and **customer retention modelling**, reducing financial risk and improving investment decision workflows by 25% for Fortune 500 clients.
- Architected Generative AI systems using **LangGraph, LangChain, RAG,** and **LLM fine-tuning**, elevating domain-specific response quality by 40% in cybersecurity and compliance auditing use cases.
- Streamlined data pipelines with **Apache Spark, Kafka, Airflow,** and **AWS**, processing over 80M+ transactions daily, automating ingestion, feature engineering, and behavioral drift analysis for model retraining.
- Operationalized AI models in cloud-native environments using **Docker, Kubernetes, AWS SageMaker AI,** and **FastAPI**, achieving sub-200ms latency.
- Systematized workflows with **MLflow, GitHub Actions,** and **CI/CD** pipelines, supporting continuous model retraining, drift detection, and versioned deployments with 99.5% deployment success rate.
- Proficient in AI/ML stacks, including **PyTorch, TensorFlow, Hugging Face**, Scikit-learn, BERT, LLaMA, GPT, prompt engineering, and agentic AI systems, in regulated enterprise contexts.

SKILLS

Languages	Python, SQL, JavaScript
ML Frameworks & Libraries	PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, Random Forest, Hugging Face Transformers
Generative AI & NLP	GPT-4, LLaMA, BERT, QLoRA, LangGraph, LangChain, Retrieval-Augmented Generation, Prompt Engineering, Tokenization, Few-shot Learning, CoT Reasoning, Vector Databases, Named Entity Recognition (NER), Sentiment Analysis
MLOps & Deployment	MLflow, Docker, Kubernetes, AWS SageMaker, FastAPI, GitHub Actions, CI/CD pipelines, Canary Deployments, A/B Testing
Data Engineering & Big Data	Apache Spark, Kafka, Airflow, Feature Engineering, Data Cleaning, Normalization
Cloud Platforms	AWS (SageMaker, Lambda, EC2, S3), GCP (VertexAI), Azure (Databricks)
Model Development	Supervised & Unsupervised Learning, LLM Finetuning, Reinforcement Learning
Model Evaluation	SHAP, ROC-AUC, F1 Score, Precision, Recall, Cross-Validation, Statistical Testing, BLEU, ROUGE
Data Science	Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly
Software Development	OOP Principles, Multi-threading, Design Patterns, Agile/Scrum Methodologies, Version Control (Git, GitHub)

EDUCATION

MASTER OF SCIENCE ***in Artificial Intelligence and Robotics***, Arizona State University, Tempe, AZ
BACHELOR OF ENGINEERING ***in Electronics Engineering***, Mumbai University, India

EXPERIENCE

FIDELITY INVESTMENTS | AI ENGINEER, San Francisco, CA **Aug 2024 - Present**

- Built **GenAI pipeline** automating Chief Investment Office workflows, integrating portfolio optimization, real-time market intelligence, and compliance monitoring, serving 500+ investment professionals
- Constructed predictive portfolio allocation models using **XGBoost** and **Scikit-learn**, incorporating historical returns, and improving risk-adjusted asset allocation strategies by 12% across live trading.
- Orchestrated data pipelines using **Apache Kafka**, **AWS Glue**, and **Athena**, processing market feeds, financial news, and regulatory updates with 99.8% into structured event streams for downstream GenAI agents.
- Formulated **feature engineering** framework for financial signals, generating time-series windows, cross-sector risk indicators, supporting both traditional ML models and LLM decision agents.

- Executed **multi-agent orchestration** with **LangGraph**, combining research, compliance, and reporting agents, with **semantic routing** and task-specific prompt engineering, reducing manual memo creation time by 50%.
- Optimized large language models (FinGPT, Palmyra) using **QLoRA** and **PEFT (Parameter Efficient Fine-Tuning)**, optimizing financial summarization and document classification while reducing GPU memory by 60%.
- Crafted **prompt engineering** system with **few-shot learning, chain-of-thought** prompting, and context construction for personalized market analysis generation and client-ready investment narratives.
- Established a **RAG** (Retrieval-Augmented Generation) pipeline leveraging **FAISS** and **Redis** for indexing of earnings reports, analyst notes, and regulatory filings, achieving a 35% improvement in retrieval relevance.
- Launched LLM systems and ML models using **AWS SageMaker**, and **Docker**, establishing real-time endpoints with **CI/CD**, reducing deployment cycle time by 30% and ensuring version control.
- Instituted **MLOps** and LLMOps infrastructure with **MLflow**, supporting automated model retraining, continual evaluation, drift detection, and reproducible experiment tracking in production environments.
- Pioneered the development of **LLM evaluations**, implementing metrics like **BLEU, ROUGE, perplexity,** and financial relevance scoring, enabling consistent quality assessment for generative outputs and agent responses.

INTERSOURCES | AI ENGINEER, Mesa, AZ **May 2024 - Aug 2024**

- Engineered a **cybersecurity auditing assistant** integrating ISO/Soc2/GRC frameworks with company-specific evidence, enabling auditors to resolve 60% of compliance queries autonomously via a conversational AI interface.
- Deployed hybrid **ML-LLM architecture**, combining XGBoost-based risk scoring and a LLaMA-powered RAG pipeline, cutting average audit preparation time by **34%** and reducing external audit findings by **57%**.
- Constructed **data pipelines** with **Kafka, Airflow, and Spark** to ingest logs, IAM exports, vulnerability scans, and policy documents, automating feature generation for both real-time and batch inference workflows.
- Implemented **evaluation framework** with retrieval metrics (Precision@k, MRR) and LLM scoring (factual accuracy, completeness), lowering incorrect citations by **18%** and inference costs by **11%**.
- Instituted **FastAPI** and **SageMaker endpoints** for low-latency interaction, enabling auditors to access control risk scores, flagged anomalies, and remediation suggestions in real time.
- Monitored MLflow lifecycle management, versioning models, prompts, and evaluation reports, while integrating Grafana-based monitoring to trigger retraining.

CAPGEMINI | AI/ML ENGINEER, Mumbai, India **May 2020 - Dec 2022**

- Delivered a fraud detection system using **PyTorch** and deep neural networks, reducing false positives by **31%** while improving anomaly detection sensitivity in high-volume transactional environments.
- Calibrated **classification models** resolving severe class imbalance via **SMOTE** and financial KPI incorporation, boosting recall across 5+ financial institutions' data.
- Synthesized **NLP** pipelines with Hugging Face Transformers **(BERT)** and custom tokenization to extract risk entities, intent signals**,** and **customer sentiment** from KYC documents, improving monitoring time by 30%.
- Containerized ML inference services with **Docker**, and **AWS SageMaker** Endpoints, achieving a 38% reduction in model deployment time and enabling **A/B testing** and **canary rollouts** in production.
- Integrated fraud and customer behavior scoring using **FastAPI**, integrating real-time predictions for both **customer NPS segmentation** (promoter, passive, detractor) and fraud detection with sub-200ms latency on **Lambda**.
- Accelerated hyperparameter optimization using **Optuna**, enhancing model convergence speed and improving F1-score compared to random by 10%, avoiding overfitting legacy patterns.
- Visualized explainable AI dashboards with **Plotly** and Seaborn, integrating **SHAP** values, feature attributions, and fraud signal analytics, reducing compliance investigation time by 19% and enhancing audit trail transparency.
- Standardized the ML lifecycle with MLflow and **GitHub Actions**, for model retraining, **data versioning**, and **drift detection**, cutting manual intervention by 25% and ensuring deployments across staging and production.
- Validated model pipelines using **ROC-AUC**, **cross-validation**, and statistical tests with **Statsmodels**, generating regulatory-grade risk scores with high interpretability, meeting strict compliance and audit standards.