Mayank Patel
Generative AI & Data Engineer

Toronto, ON | Phone: +1 289-505-3807| | Mail: 1mayank.patel0@gmail.com

---

PROFESSIONAL SUMMARY

Generative AI & Data Engineering Specialist with around 6 years of experience designing, deploying, and scaling AI/ML solutions across financial services, insurance, and enterprise technology sectors. Proven track record in LLM fine-tuning, RAG pipeline optimization, and multi-modal AI system development, delivering up to 30% latency reduction and 25% accuracy improvement. Skilled across AWS, Azure, GCP, and advanced MLOps practices, ensuring enterprise-grade scalability, compliance (PIPEDA, PHIPA, SOC 2), and measurable business outcomes. Successfully deployed AI solutions that process 10M+ customer queries annually, enhancing resolution speed, driving revenue growth, achieving significant cost savings, and improving user satisfaction.

TECHNICAL SKILLS

- **Generative AI & LLMs:** Azure OpenAI Service, AWS Bedrock, Google Vertex AI, OpenAI API, Anthropic Claude, Hugging Face Transformers, LangChain, LlamaIndex, LangGraph, GPT-4, GPT-3.5, Falcon, MPT, Mistral, Stable Diffusion, DALLE, Whisper, Bloom, Prompt Engineering, Fine-tuning, Instruction Tuning, Low-Rank Adaptation (LoRA), Reinforcement Learning from Human Feedback (RLHF), RAG, Vector Databases (Pinecone, Weaviate, Chroma, FAISS), Embedding Models (OpenAI Ada, Sentence-BERT), Multi-modal AI, AI Agents, Tool Calling, Function Calling
- **Programming & Frameworks:** Python, SQL, PySpark, Java, Scala, JavaScript, TypeScript, FastAPI, Flask, Streamlit, Gradio
- **Data Engineering & Pipelines:** Databricks, Snowflake, Apache Spark, Airflow, Kafka, Delta Lake, ETL/ELT Pipelines, Data Lakehouse, Feature Stores, Data Hub
- **MLOps & DevOps:** MLflow, Kubeflow, AWS SageMaker, Azure Machine Learning, CI/CD (Jenkins, GitHub Actions, GitLab CI), Docker, Kubernetes, Terraform, Helm
- **Cloud Platforms:** AWS (IAM, EC2, S3, Lambda, EMR, CloudWatch, SageMaker, DynamoDB, SNS, SQS, DLQ, QuickSight), Azure (Functions, Databricks, Data Factory, Synapse, Cosmos DB), GCP (Vertex AI, BigQuery, Dataflow, Pub/Sub)
- **Monitoring & Incident Management:** Splunk, Datadog, New Relic, PagerDuty, CloudWatch, Prometheus, Grafana, ServiceNow
- **Visualization & Reporting:** Power BI, AWS QuickSight, Tableau
- **Compliance & Security:** PIPEDA, PHIPA, SOC 2, Responsible AI Practices, Bias Mitigation, Model Explainability

---

PROFESSIONAL EXPERIENCE

Principal Associate Software Engineer
Capital One, Toronto, ON | Oct 2024 – Present

- Designed, developed, and deployed multi-modal AI solutions (text, image, speech) using Azure OpenAI, AWS Bedrock, and Google Vertex AI for conversational AI, summarization, and enterprise automation.
- Implemented RAG pipelines with LangChain, LlamaIndex, and vector databases (Pinecone, Weaviate, FAISS), enhancing semantic search and contextual retrieval accuracy by 25%.
- Fine-tuned LLMs (GPT-4, Mistral, Falcon) with LoRA & RLHF, aligning outputs to regulatory, compliance, and business-specific requirements.
- Developed and integrated AI-powered microservices using FastAPI, Flask, and Streamlit into enterprise applications for customer support, fraud detection, and document processing.
- Engineered large-scale ETL/ELT data pipelines in Databricks and Spark to support training and inference workloads with billions of records.
- Established MLOps pipelines with MLflow, Kubeflow, Jenkins, and GitHub Actions for continuous integration, automated testing, and one-click deployments.
- Deployed AI workloads across AWS, Azure, and GCP, implementing cost optimization strategies to reduce operational expenses by 15%.
- Collaborated with cross-functional teams — including data scientists, cloud engineers, compliance officers, and business stakeholders — to align AI solutions with product roadmaps and regulatory requirements.
- Ensure Responsible AI compliance including bias mitigation and alignment with PIPEDA, PHIPA, SOC 2.
- Led cross-functional proof-of-concept (POC) initiatives exploring emerging generative AI technologies, successfully validating new product features that influenced strategic AI investments and were adopted in production.

- Mentored junior engineers and data scientists in best practices for LLM fine-tuning, MLOps, and scalable AI deployment, resulting in faster onboarding and code quality.

Key Achievements:

- Reduced LLM inference latency by 30% through optimized RAG pipelines, improving customer query resolution speed and enhancing user satisfaction scores by 15%.
- Increased semantic search accuracy by 25% using vector embeddings and fine-tuned model.
- Delivered multi-cloud AI infrastructure enabling 99.9% uptime for production AI services.

Data Scientist / Generative AI

Intact, Toronto, ON | Dec 2021 – Sep 2024

- Designed and deployed enterprise-grade generative AI applications leveraging Azure OpenAI, AWS Bedrock, Vertex AI, and Hugging Face Transformers for internal and customer-facing use cases.
- Built multi-modal AI pipelines integrating DALLE, Stable Diffusion, Whisper, and Bloom for automated content generation, image recognition, and audio transcription.
- Created RAG-enabled search systems with LangChain, LlamaIndex, and vector databases (Pinecone, Chroma, FAISS) to improve search relevance and contextual Q&A accuracy.
- Engineered data pipelines using Databricks, Apache Spark, Delta Lake, and Kafka for high-volume, low-latency AI model training and inference.
- Implemented MLOps workflows for model lifecycle management, automated retraining, and continuous monitoring using MLflow, Kubeflow, Jenkins, and GitHub Actions.
- Built and deployed AI microservices using FastAPI, Flask, and Gradio, enabling seamless integration into enterprise applications.
- Ensured adherence to Responsible AI standards, applying bias detection tools, explainability frameworks, and regulatory compliance for all deployed models.
- Worked closely with cross-functional teams, including product managers, business analysts, software engineers, and compliance teams to define AI requirements, prioritize features, and ensure smooth delivery.

Key Achievements:

- Increased LLM response accuracy by 20% via advanced instruction tuning and fine-tuning pipelines.
- Reduced AI deployment time by 40% via automated MLOps pipelines, accelerating time-to-market for new AI-powered features.
- Deployed AI solutions that handled 10M+ customer queries annually with improved resolution speed.

Data Scientist

SOTI, India | May 2020 – Nov 2021

- Developed and deployed machine learning models for predictive analytics, anomaly detection, and operational optimization using Python, PySpark, and SQL.
- Engineered features and preprocessing pipelines for structured and semi-structured datasets to improve model accuracy and reliability.
- Conducted exploratory data analysis, statistical modeling, and hypothesis testing to derive actionable insights.
- Built ETL workflows and data pipelines for ML training and inference.
- Implemented model evaluation, hyperparameter tuning, and performance monitoring using Scikit-learn, TensorFlow, PyTorch, and MLflow.
- Designed dashboards using Power BI, Tableau, and Python libraries (Matplotlib, Seaborn).
- Applied Responsible AI practices including bias detection and adherence to compliance standards.

Key Achievements:

- Improved predictive model accuracy by 20% through advanced feature engineering.
- Automated data preprocessing pipelines, reducing manual effort by 30%.
- Delivered actionable business insights through dashboards and predictive analytics.

Data Engineer
Veeva Systems, India | May 2019 – Apr 2020

- Designed, developed, and maintained ETL pipelines for structured and semi-structured data from SQL and NoSQL databases.
- Built data warehouse and data lake architectures for analytics and machine learning initiatives.
- Implemented batch and streaming data pipelines using Apache Spark, Hadoop MapReduce, Kafka, and Flume.
- Optimized data ingestion workflows to reduce latency and improve throughput.
- Worked on feature extraction, data validation, and cleansing for predictive analytics.
- Managed data versioning, lineage, and governance for compliance.
- Created dashboards and reports using Tableau, Power BI, and SQL queries.

Key Achievements:

- Improved ETL pipeline performance by 30% by optimizing Spark jobs.
- Reduced data latency for reporting dashboards by 25%.
- Implemented data quality checks, increasing analytics accuracy by 20%.

EDUCATION

Bachelor of Technology in Computer Science — India | May 2019