

Aditya Sankranthi — AI/ML Engineer

(530) 562-5883 | sankranthiadi@gmail.com | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY:

- Around 4 years of experience in Artificial Intelligence and Machine Learning, designing, developing, and implementing machine learning solutions across retail, financial services, and healthcare domains.
- Strong expertise in machine learning, deep learning, predictive analytics, natural language processing, generative AI, and statistical modeling for solving business and enterprise problems.
- Proficient in programming languages including Python, SQL, and Java for machine learning development, data analysis, automation, and AI application development.
- Experienced in building and optimizing machine learning models including classification, regression, recommendation systems, ranking models, NLP solutions, and time-series forecasting models.
- Hands-on experience with deep learning architectures, transformer-based models, large language models (LLMs), retrieval-augmented generation (RAG), and AI-driven knowledge retrieval systems.
- Skilled in data processing, exploratory data analysis, feature engineering, and preparing structured and unstructured datasets for machine learning and AI workflows.
- Experienced in developing enterprise AI solutions including internal AI assistants, GPT-based pilots, document intelligence solutions, and generative AI applications.
- Proficient in model evaluation and optimization using cross-validation, hyperparameter tuning, performance metrics, error analysis, and model improvement techniques.
- Experienced with MLOps practices including experiment tracking, model versioning, deployment workflows, monitoring, and lifecycle management of machine learning models.
- Experienced with cloud technologies across GCP, Microsoft Azure, and AWS, including Google Cloud Storage, BigQuery, Vertex AI, Azure Blob Storage, Azure Machine Learning, Azure Cognitive Search, Amazon S3, Amazon SageMaker, AWS Lambda, Amazon EKS, and Amazon Bedrock for data processing, ML experimentation, model deployment, and enterprise AI solutions.
- Familiar with cloud-based AI platforms including AWS SageMaker for model training, experimentation, deployment, and scalable AI solution development.
- Experienced in developing LLM applications using prompt engineering, embedding models, vector search, knowledge retrieval, and RAG-based architectures.
- Knowledgeable in AI governance practices including model explainability, responsible AI principles, data privacy, and bias evaluation.
- Experienced working in Agile/Scrum environments with cross-functional teams including data engineers, software engineers, product owners, and business stakeholders.
- Skilled in creating technical documentation including model design documents, experiment reports, architecture documentation, and AI solution guidelines.
- Strong problem-solving, analytical thinking, communication, and collaboration skills with the ability to translate business requirements into scalable AI/ML solutions.

CERTIFICATIONS:

- AWS Certified Solutions Architect – Associate March 2026 - March 2029
- Claud 101 - Anthropic AI Fluency: Framework and Foundations - Anthropic

EDUCATION:

Master of Science in Computer Science @ University of Wisconsin- Milwaukee

WORK EXPERIENCE:

AI Engineer @ HCA Healthcare | Nashville, TN

Jul 2025 – Present

- Designed and developed enterprise generative AI solutions for healthcare use cases, focusing on clinical knowledge retrieval, AI assistants, and intelligent workflow automation.
- Built retrieval-augmented generation (RAG) pipelines by integrating large language models with healthcare document repositories to improve contextual information retrieval.

- Developed document processing workflows for extracting, chunking, embedding, and indexing healthcare content for enterprise AI applications.
- Implemented embedding-based semantic search and vector retrieval solutions to enable accurate knowledge discovery across large healthcare datasets.
- Designed agentic AI workflows using LLM-based reasoning, tool integration, and automated task execution approaches for healthcare business processes.
- Developed prompt engineering strategies, response evaluation methods, and optimization techniques to improve LLM output quality, relevance, and reliability.
- Integrated enterprise generative AI solutions with AWS services, using Amazon S3 for document and data storage and Amazon SageMaker for machine learning experimentation and model workflows.
- Leveraged Amazon Bedrock capabilities for enterprise LLM experimentation and integration within secure generative AI workflows.
- Used AWS Lambda to support event-driven processing and automation components within AI application workflows.
- Developed containerized AI services using Docker and deployed scalable workloads through Amazon EKS/Kubernetes for enterprise application integration.
- Integrated AI workflows with secure data pipelines while considering healthcare data privacy, compliance requirements, and responsible AI practices.
- Performed model and RAG evaluation using retrieval quality analysis, response accuracy checks, relevance scoring, and user feedback analysis.
- Implemented LLMOps practices including AI workflow versioning, experiment tracking, monitoring, evaluation, and continuous improvement of generative AI applications.
- Utilized MLflow for tracking AI experiments, documenting model iterations, comparing evaluation results, and maintaining reproducibility across development workflows.
- Monitored AI application performance, retrieval effectiveness, and system behavior to identify improvement opportunities and support continuous optimization.
- Collaborated with healthcare stakeholders, data engineers, software developers, and product teams to translate business requirements into AI-powered solutions.
- Participated in Agile ceremonies including sprint planning, technical reviews, backlog refinement, and cross-functional collaboration.
- Created technical documentation including AI architecture designs, RAG workflow documentation, prompt guidelines, evaluation reports, and deployment procedures.

Technologies Used: Python, SQL, PyTorch, LangChain, Large Language Models (LLMs), RAG, Vector Databases, AWS, Amazon S3, Amazon SageMaker, AWS Lambda, Amazon EKS, Amazon Bedrock, MLflow, Docker, Kubernetes, JIRA.

AI/ML Engineer @ Citigroup | New York, NY

Aug 2022 – Jul 2023

- Developed AI solutions for financial services use cases, focusing on early enterprise adoption of large language models, internal AI assistants, and knowledge retrieval applications.
- Worked with business teams to identify opportunities for applying generative AI technologies to improve employee productivity and information access.
- Prepared and processed structured and unstructured enterprise data including documents, reports, policies, and knowledge repositories for LLM evaluation and NLP workflows, leveraging Azure Blob Storage for cloud-based data management.
- Developed NLP pipelines for document understanding, text processing, semantic search, and enterprise knowledge discovery using Python and transformer-based NLP techniques.
- Built GPT-based AI pilots using prompt engineering techniques to evaluate conversational AI capabilities for internal business workflows.
- Developed semantic document retrieval solutions using embedding models and Azure Cognitive Search to improve internal knowledge search and information access.
- Designed prototype knowledge retrieval workflows combining semantic search and document indexing for internal financial documents.
- Leveraged Azure Machine Learning for AI experimentation, model evaluation, workflow development, and tracking NLP/LLM evaluation activities.

- Performed prompt testing, response evaluation, and optimization to improve the quality, relevance, and consistency of LLM-generated responses.
- Applied responsible AI practices including response validation, data privacy considerations, and evaluation of model behavior.
- Used MLflow for tracking experiments, comparing AI workflows, and documenting model evaluation results.
- Developed prototype LLM evaluation and NLP workflows using Docker-based environments, Azure cloud resources, and internal data repositories for demonstrations and validation.
- Collaborated with data engineers, analysts, software developers, and business stakeholders to refine AI use cases and define technical requirements.
- Participated in Agile development processes including sprint planning, requirement discussions, technical reviews, and iterative AI experimentation.
- Created documentation for AI prototypes, prompt strategies, retrieval workflows, Azure-based AI components, and evaluation methodologies.
- Supported enterprise AI adoption initiatives by researching emerging LLM capabilities and identifying practical business applications.

Technologies Used: Python, SQL, PyTorch, Hugging Face Transformers, OpenAI GPT Models, NLP, Sentence Transformers, Microsoft Azure, Azure Blob Storage, Azure Machine Learning, Azure Cognitive Search, Azure Functions, MLflow, Docker, Elasticsearch, JIRA.

Machine Learning Engineer @ Target | Minneapolis, MN

Apr 2020 – Jul 2022

- Developed traditional machine learning solutions for retail analytics, customer behavior modeling, product recommendations, ranking, and personalization use cases.
- Analyzed customer, product, transaction, and browsing datasets stored in Google BigQuery and Google Cloud Storage to identify purchasing patterns and generate insights for retail decision-making.
- Performed data preprocessing activities including data cleaning, missing value treatment, categorical encoding, feature scaling, and data transformation for machine learning pipelines.
- Designed and engineered customer, product, and transaction-level features to improve recommendation systems, predictive analytics models, and personalized retail experiences.
- Built supervised machine learning models for customer segmentation, purchase prediction, product affinity analysis, and retail forecasting applications using Scikit-learn.
- Developed recommendation system solutions using collaborative filtering and similarity-based approaches to improve personalized product discovery.
- Implemented ranking models to prioritize relevant products based on customer preferences, historical interactions, browsing activity, and engagement behavior.
- Applied deep learning techniques using TensorFlow for customer behavior analysis and complex pattern recognition across large retail datasets.
- Used GCP Vertex AI for machine learning experimentation, model training workflows, and evaluation of predictive models.
- Leveraged BigQuery for analytical data processing and Google Cloud Storage for storing machine learning datasets and intermediate artifacts.
- Evaluated recommendation and machine learning models using precision, recall, F1-score, RMSE, ranking metrics, and validation techniques.
- Performed feature optimization, hyperparameter tuning, and model experimentation to improve model performance and reliability.
- Used MLflow for experiment tracking, model comparison, documentation, and maintaining reproducible machine learning workflows.
- Collaborated with data engineers, business analysts, and product teams to integrate machine learning solutions into retail analytics processes.
- Participated in Agile ceremonies including sprint planning, backlog refinement, technical discussions, and code reviews.

Technologies Used: Python, SQL, Scikit-learn, TensorFlow, Pandas, NumPy, Google Cloud Platform (GCP), Google Cloud Storage (GCS), BigQuery, Vertex AI, Cloud Functions, MLflow, JIRA.

PROJECTS:

AI Research Agent — Agentic RAG with Verified Citations | -

- Built an agentic RAG research agent from scratch with no orchestration framework — plans coverage, searches the web, reads sources, and reflects with dynamic early-stop to produce a single cited report with mechanically verified citations.
- Achieved 91.1% citation accuracy — highest on the DeepResearch Bench public leaderboard (FACT metric) — running on a small, low-cost model (claude-haiku-4-5), outperforming dedicated deep-research products from Anthropic, OpenAI, Google, and Perplexity on citation accuracy.
- Designed a deterministic citation-grounding layer that verifies every cited URL against gathered sources and flags unverifiable claims, making citation trust a mechanical guarantee rather than a model behavior.
- Engineered a provider-agnostic model gateway (litellm) to run the same agent locally (Ollama) or via hosted APIs, plus a reusable Tool abstraction for pluggable search and page-fetch backends.
- Added a coverage planner with atomic-evidence storage and write-once synthesis, doubling valid citations per report (~16 → ~33) and lifting RACE overall into the leaderboard's top-10 band.
- Shipped as a single Dockerized service (FastAPI + SSE streaming, React/Vite/Tailwind UI) with an evaluation harness using an independent LLM judge, 100+ network-mocked tests, and GitHub Actions CI.

TECHNICAL SKILLS:

- **Programming Languages:** Python, SQL, Java.
- **Machine Learning Frameworks:** Scikit-learn, TensorFlow, PyTorch.
- **Deep Learning Architectures:** Neural Networks, CNN, Transformers, Autoencoders, Recurrent Neural Networks.
- **Generative AI & LLM Technologies:** Large Language Models (LLMs), GPT Models, Prompt Engineering, Retrieval-Augmented Generation (RAG), Embeddings, Vector Search, AI Assistants.
- **Machine Learning Techniques:** Classification, Regression, Clustering, Recommendation Systems, Ranking Models, Personalization, Predictive Modeling, Feature Selection.
- **Natural Language Processing:** Text Classification, Text Preprocessing, Sentiment Analysis, Named Entity Recognition, Semantic Search, Document Intelligence.
- **Data Processing & Feature Engineering:** Pandas, NumPy, Data Cleaning, Data Transformation, Exploratory Data Analysis, Feature Engineering, Data Validation.
- **MLOps Tools:** MLflow, Docker, CI/CD Pipelines, Model Versioning, Model Monitoring, Experiment Tracking.
- **Cloud Platforms - GCP:** Google Cloud Platform (GCP), Google Cloud Storage (GCS), BigQuery, Vertex AI, Cloud Run, Cloud Functions.
- **Cloud Platforms - Microsoft Azure:** Microsoft Azure, Azure Blob Storage, Azure Machine Learning, Azure Cognitive Search, Azure Functions, Azure App Service.
- **Cloud Platforms - AWS:** Amazon Web Services (AWS), Amazon S3, Amazon SageMaker, AWS Lambda, Amazon EKS, Amazon Bedrock.
- **Databases:** SQL Server, PostgreSQL, Oracle, MongoDB, Data Warehousing.
- **Model Evaluation & Optimization:** Cross-Validation, Hyperparameter Tuning, Grid Search, Random Search, Precision, Recall, F1-Score, ROC-AUC, RMSE.
- **AI Deployment:** REST APIs, Model Serving, Containerized Deployments, Enterprise AI Integration.
- **Data Visualization:** Matplotlib, Seaborn, Tableau.
- **Development Methodologies:** Agile, Scrum, SDLC, JIRA, Version Control Practices.
- **AI Governance:** Model Explainability, Responsible AI, Bias Detection, Data Privacy, Compliance Awareness.
- **Recommendation & Search Systems:** Collaborative Filtering, Content-Based Recommendations, Ranking Algorithms, Personalization Models.
- **Enterprise AI Concepts:** Agentic Workflows, Knowledge Retrieval Systems, AI Automation, LLMops, AI Platform Integration.