Himakar Yanamandra

San Francisco, CA | (412) 284 3962 | hyanaman@tepper.cmu.edu | LinkedIn

SUMMARY

Applied AI Engineer with 6+ years building production ML and AI agent systems. Proven expertise in LLM API integration, workflow automation, and real-time monitoring of AI agents in production environments.

WORK EXPERIENCE

PROJECTS YARD AI **Jan 2026 - Present**

Senior Machine Learning Engineer

- Built RAG pipeline using embedding models and vector search, reducing user portfolio creation time from 2 hours to 15 minutes through prompt engineering and retrieval optimization.
- Developed end-to-end AI agents leveraging LLM APIs for automated case study generation, integrating prompt construction, response handling, and tool use in production Python workflows.
- Integrated AI tools with external platforms using APIs (Slack, HubSpot), handled authentication, rate limits, error cases, and logging for seamless workflow automation.
- Monitored and maintained deployed AI agents in production, tracked quality metrics, triaged failures, and shipped improvements based on real usage data.
- Built evaluation harnesses to catch regressions and measure agent quality programmatically, incorporating human feedback loops for continuous improvement.

HONDA RESEARCH INSTITUTE (Capstone Project) **Aug 2025 - Dec 2025**

Senior Machine Learning Engineer Intern

- Designed a multi-agent conversational AI system using LLMs and RAG to automate trend research across unstructured sources, reducing manual analysis time by 30% and delivering ranked insights with executive-ready briefs.
- Built model evaluation framework (precision, relevance, latency metrics) with human-in-the-loop feedback loops to improve output.
- Monitored deployed agents in production, tracked quality metrics, triaged failures, and shipped improvements based on real usage data.

ASWARTHA TECHNICAL SERVICES **Jan 2024 - Dec 2024**

Senior Machine Learning Engineer

- Designed and deployed data lake architecture on AWS (S3, Glue, Lambda) across multiple regions, enabling centralized data access for ML pipelines which reduced downtime by 15%.
- Built end-to-end ML pipelines for predictive analytics dashboards, implementing feature engineering, model training, and real-time inference on AWS SageMaker with automated model monitoring.
- Connected ML and analytics pipelines to Snowflake via API for unified access, implemented logs and error handling to ensure quality and reliability.

CREDIT SAISON **Dec 2022 - Dec 2023**

Senior Machine Learning Engineer

- Architected alternative credit scoring system using ML models and deep learning on behavioral proxy data, unlocking credit access for 100K+ underserved businesses and driving $500M+ in disbursals in year one.
- Launched a real-time ML decisioning service for fraud and credit risk scoring, improving approval turnaround by 4x across 1.2M active loans ($1.2B AUM) on AWS SageMaker with sub-100ms latency.
- Shipped 7 lending partner integrations (Amazon, Airtel, Booking.com) by collaborating across cross-functional teams (product, engineering, data science), reducing development cycle from 20 days to 3 days and cutting infrastructure costs by 36%.

MICRON **Jul 2021 - Nov 2022**

Senior Machine Learning Engineer

- Deployed clustering and anomaly detection models on high-volume test data (TBs daily), built data pipelines for failure signature identification, reducing debug cycle time by 18%.

LANGUAGE TRANSLATIONS RESEARCH CENTRE, IIIT-H **Jan 2018 - Jun 2021**

Machine Learning Engineer

- Published 3 peer-reviewed papers on healthcare NLP and sentiment analysis. Led 40+ member research program building 25 medical NLP datasets using transformer architectures (BERT, seq2seq) for classification and multilingual translation.

EDUCATION

CARNEGIE MELLON UNIVERSITY **Jan 2025 - Dec 2025**

Master of Science (M.S.), Product Management (Tech MBA) *Pittsburgh, PA*

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, HYDERABAD **Jul 2019 - Jul 2022**

Master of Science (M.S.), Computer Science and Engineering *Hyderabad, India*

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, HYDERABAD **Jul 2015 - Jul 2019**

Bachelor of Technology (B. Tech), Computer Science and Engineering *Hyderabad, India*

SKILLS

- **Platform Integrations**: Slack API, HubSpot API, Jira API
- **MLOps**: Production Python, Model Deployment, ML Pipelines, Model Evaluation, API Integration, Real-Time Serving, A/B Testing, Docker, Airflow
- **Languages/Tools**: Python, SQL, Pandas, NumPy, Spark
- **ML/AI**: PyTorch, TensorFlow, scikit-learn, XGBoost, LLMs, RAG, NLP/NLU, Conversational AI

- **Cloud / Data**: AWS (SageMaker, EC2, S3, Glue, Lambda), GCP (Vertex AI, BigQuery), Azure, Snowflake, PostgreSQL, MongoDB, Kubernetes, Distributed Systems
- **AI Agent Engineering**: LLM APIs, Prompt Engineering, Response Handling, Context Management, Tool Use, Workflow Automation

RESEARCH PUBLICATIONS

- Smokeng: Towards Fine-Grained Classification of Tobacco-Related Social Media Text.*W-NUT, EMNLP 2019*
- Smokpro: Towards Tobacco Product Identification in Social Media Text.*SIIRH Workshop, ECIR 2020*
- Towards Sentiment Analysis of Tobacco Products' Usage in Social Media.*RANLP 2021*