Hema Harsha Vardhan Peela

AI/ML Engineer

Hpeela5@gmail.com | 806-544-0868| Austin, TX| LinkedIn| GitHub | Portfolio

SUMMARY

AI/ML Engineer with 4+ years across data and AI engineering, including 1.5+ years building LLM-powered AI agents, multi-agent systems, and production RAG pipelines in Python, grounded in large-scale data engineering. Currently technical lead for two AI agent products at Johnson & Johnson, built on AWS and Azure OpenAI; previously served 5M+ monthly LLM inferences at Wells Fargo with an evaluation framework that reduced hallucinations by ~30%.

PROFESSIONAL EXPERIENCE

AI/ML Engineer | Johnson & Johnson **Aug '26 – Present**

JJ Customer Connect AI Agent

- Led a four-person team as technical lead to build a Langflow and FastAPI assistant for J&J B2B order, invoice, and product queries, passing 24 of 25 development test cases with approximately four-second responses.
- Integrated eight live J&J APIs through a router that directs questions to semantic or analytical processing, grounding responses in real-time customer data.
- Independently deployed the custom backend and Langflow workflow on AWS Bedrock AgentCore; implemented OAuth 2.0 token-based authorization (verified against the client's database) and user-scoped conversation storage in AWS RDS PostgreSQL.

SAVIA – Multi-Agent SDLC Assistant

- Led a five-person team as technical lead and built the Python, FastAPI, and Azure OpenAI backend for an assistant that processes requirements from documents, recordings, and text to recommend SDLC agents.
- Designed specialist SDLC agents around a shared base class, registry, and unified API, enabling new agents to be added with one implementation file and one registry entry.
- Grounded generated outputs in the source input, removing invented names, dates, and figures, and calculated test coverage from story-to-test mappings, backed by 38 backend tests and sample inputs for every agent.
- Implemented review and approval workflows with section-level edit suggestions, open-question resolution, recorded approver sign-off, and Word, Excel, and PowerPoint exports.

AI/ML Engineer | Wells Fargo **May '25 – Jul '26**

- Served 5M+ LLM inferences per month for customer service and cross-functional teams by building a production agentic RAG pipeline with multi-step reasoning over Pinecone vector search, OpenAI embeddings, and knowledge graphs.
- Reduced hallucinations by ~30% and caught quality regressions before release by designing an LLM evaluation framework with LLM-as-judge scoring, canary deployments, and audit-driven feedback loops.
- Shipped 7 production NLP and LLM models in a regulated environment by fine-tuning with LoRA in TensorFlow and deploying through Airflow CI/CD with PII guardrails.
- Gave RAG and analytics teams reliable data across 8 financial use cases by engineering Python and Spark pipelines that unify structured and unstructured sources, with automated data quality validation.

Data Engineer Intern | YourBook Team **Nov '24 – May '25**

- Gave stakeholders accurate KPI reporting by building 10+ Power BI and Tableau dashboards backed by SQL, Python, and R validation checks before each rollout.
- Made reporting consistent across an 8-person team and kept 50+ data quality issues out of stakeholder reports by standardizing reporting logic and auditing large datasets with senior analysts.

Data Engineer | Cognizant **Aug '21 – Jul '23**

- Raised pipeline success rate from 91% to 98%+ across 50TB+ of data by orchestrating ETL/ELT on Hadoop, Hive, and Kafka and resolving 30+ production failures per month.
- Supported 2.5M+ daily transactions by building Python, SQL, and Spark ETL workflows integrating Kafka, Hive, and Amazon Redshift.
- Eliminated ~60 hours of manual effort per month by automating PySpark, SQL, dbt, and Informatica workflows with monitoring.
- Unified 150K+ patient records per month from multiple hospitals by building AWS Glue and SQL pipelines with entity extraction and PII redaction, feeding Power BI dashboards reviewed weekly by 20+ clinical and operations leaders.

PROJECTS

NYC Taxi Demand Pipeline | PySpark, Delta Lake, Snowflake, dbt, Airflow, GitHub Actions

- Turned 9.5M raw taxi trips into analytics-ready demand data by building an end-to-end medallion (bronze/silver/gold) pipeline with PySpark and Delta Lake, loaded into Snowflake and orchestrated with Airflow.
- Eliminated duplicate rows, missing records, and wrong dates by debugging a Snowflake file-partitioning bug, and enforced data quality with 32 dbt tests run automatically through GitHub Actions CI/CD.

TECHNICAL SKILLS

- **Languages:** Python, SQL, R
- **AI/ML & LLMs:** LLMs, RAG, Agentic AI, Multi-Agent Systems, LLM Evaluation (LLM-as-judge), LoRA Fine-Tuning, NLP, PyTorch, TensorFlow, Azure OpenAI, OpenAI, Groq, Whisper, Langflow
- **Retrieval:** Pinecone, Vector Databases, Embeddings, Knowledge Graphs
- **Backend & APIs:** FastAPI, REST APIs, JWT Authentication, LDAP/Active Directory, OAuth 2.0
- **Data Engineering:** Spark/PySpark, Airflow, Kafka, Hadoop, Hive, dbt, Informatica, Delta Lake, ETL/ELT
- **Cloud & Platforms:** AWS (Bedrock AgentCore, EC2, Glue, Redshift, RDS PostgreSQL), Snowflake
- **DevOps & BI:** CI/CD, GitHub Actions, Jira, Power BI, Tableau

EDUCATION

Master of Science in Computer Science - Texas Tech University | 3.7/4.0 **May '25**