

**Sai srinivas**

Dallas – Fort Worth Area

**SUMMARY:**

Skilled AI/MLEngineer with 4 years of experience across Generative AI, NLP, and applied deep learning. Specialized in designing and fine-tuning large-scale models—LLMs, GANs, VAEs, and Diffusion models—for real-world applications in text, image, and multimodal generation. Proficient in Python, PyTorch, TensorFlow, and Hugging Face, with strong command of data preparation, model training, evaluation (BLEU, ROUGE, FID), and prompt engineering. Delivered production-grade solutions including a text-to-image generator using Stable Diffusion, a custom GPT-based chatbot for support automation, an AI writing assistant, and a voice-to-text task manager using Whisper. Skilled in deploying scalable AI systems on AWS/GCP with a focus on responsible and ethical AI.

**EXPERIENCE:**

Machine Learning Engineer.

FIRST US IT 02/2025 - Present

**Responsibilities:**

• Designed and optimized machine learning models to deliver highly accurate and efficient predictions, supporting key business objectives.
• Engineered scalable, automated data pipelines to ensure clean, high-quality input for model training and evaluation.
• Utilized deep learning and ensemble methods to improve model performance and robustness.
• Collaborated with engineering and product teams to integrate ML models into production with minimal latency.
• Applied hyperparameter tuning, cross-validation, and continuous monitoring to maintain model accuracy and prevent drift.
• Built reusable ML components and APIs for integration across multiple applications.
• Used tools like MLflow and Hugging Face to manage experiments and model deployments.
• Conducted model explainability using SHAP to interpret predictions and support business decision-making.
• Deployed ML models on AWS using SageMaker and Lambda for scalable, real-time inference.
• Collaborated in agile teams, contributing to sprint planning, code reviews, and continuous integration of ML pipelines.

**Environment:**

Python, SQL, TensorFlow, PyTorch, Scikit-learn, Docker, AWS (SageMaker, Lambda, EC2, S3), MLflow, Git, Jupyter, Pandas, NumPy, FastAPI, Bash, Matplotlib, Seaborn, VS Code, Hugging Face

 **Ai Engineer**

 **COUGAR TECH INC 05/2024 - 02/2025**

**Responsibilities:**

• Designed, developed, and deployed AI and machine learning models to address key business challenges and improve operational efficiency.
• Built scalable AI pipelines using Python, TensorFlow, and PyTorch to process and analyze large datasets effectively.
• Integrated AI models into production environments, ensuring robustness, low latency, and high availability.
• Collaborated with cross-functional teams, including product managers and engineers, to gather requirements and deliver tailored AI solutions.
• Conducted rigorous model evaluation, hyperparameter tuning, and continuous monitoring to maintain and improve model accuracy and performance.
• Implemented automated workflows for model retraining and performance tracking to adapt to changing data conditions.
• Developed custom data preprocessing and feature engineering techniques to enhance model input quality and predictive power.
• Applied advanced algorithms such as deep learning, reinforcement learning, and ensemble methods to improve solution accuracy.
• Used version control tools like Git for code management and collaborated on code reviews to maintain high-quality standards.

• Performed end-to-end UI testing, ensuring DOM perfection and responsive behavior across multiple browsers and devices, using Selenium WebDriver and browser DevTools.

• Developed and maintained Selenium WebDriver scripts for automated UI regression testing, increasing testing efficiency by 60%.

• Incorporated Selenium tests into the CI/CD pipeline, automating application flow validation after each deployment.

• Collaborated with developers and QA engineers to replicate, isolate, and resolve frontend bugs using browser tools and logs, improving product quality.

• Used Postman and browser DevTools to validate API contract adherence, error handling logic, and measure API response times.

• Delivered QA reports highlighting trends in UI test failures and slow API endpoints, contributing to continuous improvement and on-time delivery of QA support.

• Utilized browser DevTools to analyze network activity, identify performance bottlenecks, and optimize CSS/HTML and JavaScript for enhanced user experience.

• Drafted comprehensive test cases and data strategies to cover edge scenarios, ensuring robust UI functionality and meeting business requirements.

**Environment:**

Machine Learning, Deep Learning, Reinforcement Learning, Ensemble Methods, Data Preprocessing, Feature Engineering, Model Evaluation, Hyperparameter Tuning, Model Monitoring, Automated Workflows, Code Reviews, Documentation, Agile, Collaboration, Mentoring

**Prompt engineer**

**Merit american technologies 09/2022 - 10/2024**

**Responsibilities:**

Designed and optimized prompts for large language models (LLMs) to enhance tasks such as question answering, summarization, and content generation.
• Collaborated with product and engineering teams to seamlessly integrate prompt workflows into AI-driven production systems.
• Conducted evaluations and testing to improve response accuracy, minimize hallucinations, and ensure model outputs aligned with business objectives.
• Developed reusable prompt templates and authored best practice documentation to scale generative AI capabilities across multiple projects.
• Analyzed model behavior to fine-tune prompt structures for consistent, reliable, and context-aware outputs.
• Researched and implemented prompt engineering techniques to enhance LLM efficiency and user experience.
• Monitored and analyzed AI model outputs to identify and troubleshoot response inconsistencies.
• Collaborated with data scientists to incorporate feedback loops for continuous prompt refinement and model improvement.
• Developed training materials and guidelines to onboard new team members on prompt engineering best practices.
• Assisted in designing A/B tests to evaluate prompt variations and optimize user engagement metrics.

**Environment:**

Prompt Engineering, Large Language Models (LLMs), Natural Language Processing (NLP), Model Evaluation, Generative AI, Content Optimization, A/B Testing, User Experience Improvement, Documentation, Training & Onboarding, Cross-functional Collaboration, Feedback Loop Integration

**Generative ai engineer**

**Intelli Infotech 01/2022-07/2024**

**Responsibilities:**

• Supported development of AI-driven applications using large language models for content creation, summarization, and question answering.
• Assisted in designing and refining prompts to optimize model responses and overall system performance.
• Fine-tuned pre-trained models with tools such as OpenAI, Hugging Face, and LangChain to meet project requirements.
• Conducted tests and evaluations to improve output accuracy, reduce errors, and enhance model reliability.
• Collaborated with senior team members to prototype AI features and integrate them into functional applications.
• Documented experimental results and contributed to knowledge-sharing within the engineering team.
• Researched the latest advancements in generative AI to incorporate cutting-edge techniques into projects.
• Worked on debugging and troubleshooting model issues to ensure smooth application performance.
• Assisted in creating reusable prompt templates to speed up development and maintain consistency.
• Participated in team meetings and brainstorming sessions to align AI solutions with business goals.
• Supported data preprocessing and cleaning to improve input quality for model training and evaluation.
• Helped implement feedback loops to continuously refine model outputs based on user and stakeholder input

.**Environment:**

Large Language Models, Prompt Design, Prompt Optimization, Data Preprocessing, Data Cleaning, Model Evaluation, Model Testing, AI Frameworks, Problem Solving, Debugging, Collaboration, Communication, Documentation, Reporting

**SKILLS:**

• Programming languages: Python, R, Java, Bash, SQL
• Frameworks & libraries: TensorFlow, PyTorch, Hugging Face, LangChain, Scikit-learn, Keras, FastAPI, OpenAI API
• Databases: MySQL, PostgreSQL, MongoDB, Redis, Elasticsearch
• DevOps tools: Docker, Kubernetes, MLflow, Jenkins, Git, CI/CD pipelines, Apache Airflow
• Cloud platforms: AWS (SageMaker, Lambda, EC2, S3), Microsoft Azure, Google Cloud Platform (GCP), IBM Cloud
• Testing & Monitoring: Unit Testing, Integration Testing, A/B Testing, Model Validation, Cross-validation, Performance Monitoring, Automated Testing

**PROJECTS:**

**SmartPredict: Scalable ML Forecasting Engine**
**Description:**
Designed and deployed a machine learning system to generate business forecasts and intelligent insights using time-series data. Integrated ensemble learning and LSTM-based models, with continuous evaluation pipelines using MLflow. Deployed models on AWS SageMaker for real-time inference with API endpoints serving millions of requests per month.

**Impact:** Boosted forecasting accuracy by 35%, helped automate inventory planning.
**Tools:** PyTorch, TensorFlow, MLflow, SageMaker, FastAPI, Git

**Project Name:** **SupportBot-GPT: AI Customer Support Chatbot**
**Description:**
Built a production-grade chatbot for automating support queries using GPT-3.5 and LangChain. Fine-tuned prompts to handle various customer scenarios, integrated the model with FastAPI and deployed using Docker. Enabled the support team to cut first-response time by 45% and reduced ticket load by 30%.
**Tools:** LangChain, OpenAI API, FastAPI, Docker, MongoDB, Git
**Impact:** Boosted forecasting accuracy by 35%, helped automate inventory planning.
**Tools:** PyTorch, TensorFlow, MLflow, SageMaker, FastAPI, Git

**Project Name:** **SupportBot-GPT: AI Customer Support Chatbot**
**Description:**
Built a production-grade chatbot for automating support queries using GPT-3.5 and LangChain. Fine-tuned prompts to handle various customer scenarios, integrated the model with FastAPI and deployed using Docker. Enabled the support team to cut first-response time by 45% and reduced ticket load by 30%.
**Tools:** LangChain, OpenAI API, FastAPI, Docker, MongoDB, Git

**Project Name:** **PromptIQ: LLM Prompt Optimization Toolkit**
**Description:**
Developed a toolkit for optimizing prompts across LLM tasks such as summarization, question answering, and writing. Conducted prompt evaluations, structured A/B tests, and built a prompt repository with reusable templates. Integrated prompt flows into production with improved model alignment and 25% reduction in hallucinations.
**Tools:** OpenAI API, LangChain, PromptLayer, Python, Notion

**Project Name:** **GenFlow: LLM App Prototyping Suite**
**Description:**
Contributed to building an internal toolset for rapid prototyping of generative AI features. Fine-tuned models on Hugging Face, optimized prompts for creative writing and Q&A, and conducted evaluations for hallucination reduction. Helped integrate GenAI modules into early-stage product experiments.
**Tools:** Hugging Face, LangChain, Transformers, Jupyter, Streamlit

**EDUCATION:**

**• Bachelor’s in Computer Science and Engineering**

Sathyabama University 2022

Field of Study: Computer Science and Engineering

Specialization: AI&ML