**AI Machine learning Engineer Sai Srinivas**

**saisnivas9@gmail.com +1 (469)-473-6713**

**Professional Summary**

* Experienced AI professional and researcher with extensive experience in implementing projects across **Generative AI, Machine Learning, Agentic AI, Deep Learning, Computer Vision, Natural Language Processing, Predictive Analytics, and Conversational AI.**
* Currently working on designing and implementing solutions using LLM, AI Agents and driving ec2 advancements in NLP, Gen AI for the organization.
* Developed and deployed generative AI applications leveraging LLMs such as OpenAI GPT and Anthropic Claude.
* Designed and deployed scalable AI solutions using AWS Bedrock, utilizing its managed service for deploying and running generative AI models, enabling faster and more efficient model deployment.
* Developed and deployed scalable models leveraging LoRA and QLoRA for natural language processing tasks, improving accuracy and efficiency in real-world applications.
* Built and deployed machine learning models on AWS SageMaker, ensuring that models were efficiently trained and deployed for real-time inference
* Managed and optimized large datasets within Google BigQuery, leveraging its distributed architecture to ensure fast and cost-effective data analysis for AI-driven applications.
* Implemented embeddings to optimize search, classification, and semantic similarity tasks, resulting in improved performance of AI systems in knowledge retrieval and recommendation engines.
* Assisted in the design and architecture of AI systems using Gemini and other LLM frameworks, ensuring the effective translation of business requirements into technical specifications.
* Created POC solutions for leveraging generative AI models in large-scale content creation pipelines, enabling clients to automate and scale content production processes efficiently.
* Proficient in performing data preparation, data cleaning, EDA, feature engineering, data visualization, model building, parameter fine-tuning and monitoring.
* Developed and deployed Langsmith applications for seamless integration of language models, facilitating multi-modal interactions and supporting personalized AI-driven experiences.
* Implemented Guardrails for AI models to ensure compliance with privacy regulations, including automatic PII redaction, to protect sensitive information during model inference.
* Analyzed A/B testing results to make data-driven decisions, improving the chatbot's performance in real-time and boosting customer interaction efficiency.
* Employed GitHub Copilot to assist in code reviews, ensuring adherence to best practices and coding standards, and optimizing code for improved performance.
* Guided the creation of MVPs for clients in the e-commerce and advertising industries, using generative AI to produce personalized product recommendations and content based on user behavior and preferences.
* Utilized LangGraph to visualize agent behavior and logic flow, ensuring transparency and effective management of decision trees for complex LLM-based agents.
* Integrated BigQuery with various GCP services, such as Dataflow and Cloud Storage, to streamline ETL processes and enable real-time analytics for AI and business intelligence solutions.
* Developed robust workflows with AWS Step Functions, orchestrating end-to-end processes for model training, evaluation, and deployment while integrating with other AWS services like SageMaker and Lambda.
* Managed data preprocessing pipelines using Pandas and NumPy, efficiently handling large datasets for model training and inference in both structured and unstructured formats.
* Collaborated closely with product managers to define MVP requirements for Generative AI applications, ensuring quick prototyping and rapid iteration cycles.
* Worked with ERP systems to create automated processes for inventory management, order processing, and sales reporting.
* Leveraged Sourcegraph Cody to enhance code navigation, comprehension, and collaboration among development teams, resulting in improved code quality.
* Contributed to the integration of AI models into enterprise systems, particularly ERP and CRM, for improved decision-making and automation of routine tasks.
* Developed and deployed Generative AI models using TensorFlow, PyTorch, and Keras, creating high-quality text, image, and audio generation systems for various applications.
* Built and maintained Python-based AI services using LangChain and CrewAI, implementing RAG-based retrieval and Agentic AI workflows to enhance data retrieval and decision-making processes.
* Integrated OpenAI, Bard, Claude, and Azure OpenAI APIs, optimizing model performance by fine-tuning temperature, top-p, and max tokens, reducing hallucinations through embedding-based retrieval
* Hands on experience with AI Agents, LLMs like Falcon, Mistral, Llama, Bert, GPT and Gemini.
* Worked on REST API development and automation and was also involved in monitoring of developed services.
* Designed and implemented end-to-end AutoML pipelines using AutoGen, automating the model selection, hyperparameter tuning, and feature engineering processes, significantly reducing time-to-market for machine learning models.
* Designed and deployed scalable RESTful API services, securing AI endpoints using OAuth2, JWT authentication, and implementing API rate limiting to ensure system reliability and security.
* Implemented descriptive, predictive models and advanced algorithms to extract information from data sets.
* Experience in applying deep learning techniques using Tensor flow, Keras, Pytorch.
* Expertise in LLM prompt writing, prompt optimization, enhancing model accuracy and performance for a given task.

**Technical Skills:**

* **Programming:** Python, Java, JavaScript,
* **Python Packages, Libraries:** Pandas, AWS Bedrock, NumPy, Matplotlib, OpenCV, Jumpstart, Scikit-Learn, Keras, Embeddings, TensorFlow, Hugging face, FAST API.
* **Machine Learning:** Having exposure to various Supervised Learning Algorithms for regression and classification analysis, Ensemble techniques, Anomaly detection, unsupervised learning algorithms like Clustering, PCA
* **AWS Services**: AWS Lambda, AWS Step Functions, AWS CloudWatch, AWS CloudFormation, AWS Glue, SageMaker, OpenSearch Serverless
* **NLP:** RNN, LSTM, Transformers (Bert, GPT), Spacy, NLTK, Text Blob, ASR, Stanford NLP, Genism.
* **Gen AI:** Llama Index, RAG, Geming, Prompt engineering, LLM fine tuning, GitHub Copilot, Sourcegraph Cody
* **Agentic AI:** Langchain, LangGraph, Langsmith, AutoGen, Assistants API.
* **Cloud Services**: Google Vertex, Azure Open AI, BigQuery, Google Dataflow, GCP AI Services
* **Statistics:** Hypothesis testing, A/B testing, Sampling techniques
* **Development Methodologies**: Agile, Scrum, DevOps
* **Other:** Data cleaning, ERP, TTS, STT, CRM, Data pre-processing, EDA, Predictive analytics, Time series, ML model building, fine tuning and deployment, Neural networks, Model quantization, Lora, Qlora.

**Professional Experience:**

**NRG Energy (Houston, Texas) August 2024 to Till Date Role: AI Machine learning Engineer**

* Currently working with NRG Energy as AI Machine learning engineer building Agentic AI, and Gen AI based solutions using Assistants API, AutoGen, LangGraph and LLM’s.
* Developed multi-AI agent systems are designed to assist customer service representatives in solving a variety of business problems.
* Developed custom pipelines for PII redaction using AWS tools, ensuring that all AI-generated content adheres to strict data privacy standards and reduces the risk of data leakage.
* Created automated workflows for data management and preprocessing using Python, Pandas, and NumPy, streamlining data pipelines and reducing model training time.
* Utilized GitHub Copilot to assist in writing, refactoring, and optimizing code across various projects, resulting in improved development efficiency.
* Fine-tuned BERT and other LLM models using the Hugging Face Transformers library and integrated them into production systems to improve natural language understanding.
* Implemented A/B testing frameworks using Google Optimize and Optimizely, analyzing conversion rates, user interactions, and satisfaction metrics to refine AI-driven solutions.
* Developed and deployed advanced generative AI solutions using LLMs and Gemini, enhancing virtual assistants' ability to handle a variety of customer queries with high precision.
* Implemented feedback loops in MVPs, ensuring continuous improvement of Generative AI models based on user interactions and real-time data.
* Developed and deployed custom plugins and integrations for Sourcegraph Cody, enhancing its functionality and improving development workflows.
* Leveraged Langsmith for scalable AI development, ensuring smooth orchestration and deployment of language-based systems across various cloud and on-prem environments.
* Integrated Dataflow with GCP services like BigQuery and Cloud Pub/Sub to build scalable and automated data processing workflows for large-scale AI applications.
* Implemented CloudWatch for real-time monitoring and logging of AI models in production, ensuring model performance and uptime for mission-critical applications.
* Automated the deployment of AI models on AWS Lambda, enabling serverless architecture for real-time, scalable inference tasks across multiple environments.
* Integrated OpenSearch Serverless with machine learning workflows to enable real-time querying of unstructured data, supporting faster AI-driven decision-making.
* Utilized Python for data preprocessing, model training, and evaluation, ensuring seamless integration with AWS services like SageMaker, Lambda, and Step Functions to automate machine learning pipelines.
* Integrated AWS Bedrock with custom pipelines to enhance model development and reduce deployment time, ensuring alignment with business needs and technical requirements.
* Developed prototype systems for TTS and STT applications, experimenting with different architectures and fine-tuning models for improved voice recognition and synthesis.
* Designed and implemented RESTful APIs and microservices to expose AI-driven features securely, employing best practices in API security, including OAuth2, JWT, and rate limiting.
* Integrated GCP AI services, such as Vertex AI, to enable custom model deployment, training, and monitoring, driving actionable insights for the business.
* Spearheaded the development of POC projects focused on generative AI, showcasing the capabilities of AI models for tasks such as text-to-image generation and predictive text synthesis.
* Built and maintained TTS and STT systems, leveraging NLP techniques for audio/voice data manipulation and developing user-friendly AI-driven voice applications.
* Refined the RAG pipeline by fine-tuning models for both retrieval and generation tasks, enhancing accuracy and delivering more personalized content for users.
* Used FastAPI to publish developed Agentic and other services developed as REST APIs, enabling seamless interaction between intelligent agents and external systems.
* Designed and developed MVPs leveraging Generative AI to address key customer pain points and validate market demand.
* Built a comprehensive framework to benchmark and evaluate the performance of both the retrieval and generation components of RAG models, optimizing system functionality and ensuring improved overall performance.
* Built advanced NLP applications using Langgraph, enabling efficient graph-based knowledge representation and reasoning within AI models.
* Utilized Azure ML Studio to automate the end-to-end AI pipeline, from data ingestion to model evaluation and deployment, streamlining the development of custom generative AI models.
* Employed Langchain to orchestrate complex workflows, integrating external APIs and AI models to power agents with real-time, dynamic decision-making capabilities for enhanced user experiences.
* Customized and fine-tuned Jumpstart’s pre-trained models for specific business needs, enhancing performance in text generation, image synthesis, and AI-powered chatbots.
* Integrated Amazon Jumpstart with AWS SageMaker to streamline model deployment and ensure seamless scalability of AI solutions for large-scale production environments.
* Developed and optimized AI models using TensorFlow and PyTorch, focusing on RAG models to support enterprise-scale applications.
* Utilized embeddings and vector databases to improve NLP model outputs, enabling more precise answers for question-answering systems and chatbot implementations.
* Worked closely with security teams to implement secure data transmission and AI endpoint access control via OAuth2 and JWT, ensuring compliance with HIPAA regulations.
* Designed and deployed AI agents using LangGraph, CrewAI, and similar frameworks for efficient workflow management and task automation.
* Developed and deployed scalable common services using FastAPI to standardize the document vectorization process across the organization, enabling efficient handling and processing of large volumes of documents.
* Designed and developed machine learning models for classification, regression, and recommendation systems using TensorFlow, Keras, and PyTorch, focusing on both supervised and unsupervised learning tasks.
* Architected and deployed end-to-end solutions using GCP AI services, including Vertex AI, for the development and integration of advanced machine learning models into business applications.
* Developed RAG system with Assistants API using services like file search, function call, vector store to give personalized answers to the users.

**Samsung Research (Mountainview, CA) November 2021 to July 2024 Role: NLP/ GenAI Engineer**

* Worked for Samsung R&D in Bixby project which is an AI Voice Assistant for Samsung Devices.
* Work mainly involved around the building Classifiers, NER, NLG systems using various Machine Learning and Deep Learning algorithms.
* Collaborated seamlessly across teams to ensure smooth implementation of ML and Gen AI use cases for driving business values.
* Worked on commercialization of various Galaxy AI features for mobile devices.
* Responsible for model building, re-training, fine tuning, development, code maintenance, for various modules.
* Integrated chatbots with messaging platforms like Slack, WhatsApp, and Facebook Messenger, enabling seamless multi-channel user experiences.
* Designed and implemented data processing pipelines using Google Dataflow, ensuring efficient ETL processes and real-time data ingestion for AI and machine learning applications.
* Developed complex Langchain-based applications that support natural language understanding and multi-step reasoning for interactive virtual assistants and AI agents.
* Integrated Guardrails into AWS Generative AI solutions to safeguard against inappropriate outputs and enforce security measures for sensitive data.
* Prototyped new agent functionalities rapidly using Langchain, creating data pipelines and incorporating external knowledge sources to improve agent understanding and accuracy in real time
* Optimized model performance and cost-efficiency using AWS Bedrock, leveraging its high-performance infrastructure for inference tasks in large-scale applications.
* Experience in designing and implementing RAG architectures, optimizing RAG workflows, including fine-tuning retrieval systems, designing efficient vector search strategies, and improving generative models to produce contextually accurate information.
* Employed LlamaIndex to efficiently structure and index large datasets for integration with LLMs resulting in faster data retrieval and more accurate responses in NLP applications.
* Utilized Python's NLP libraries such as NLTK, spaCy, and Gensim for text preprocessing and analysis
* Improve the user and device voice interaction in Samsung smart home eco system.
* Making the developed models and modules available on various Samsung devices like mobile, watch, speaker, TV.
* Conducted user acceptance testing (UAT) for deployed chatbots, gathering feedback and refining system performance to ensure high-quality user interactions.
* Developed and deployed Agentic AI systems that combine intelligent agent-based architectures with deep learning models to solve complex decision-making problems autonomously.
* Developed and deployed generative AI models using Azure ML Studio, leveraging deep learning frameworks like TensorFlow and PyTorch to create innovative applications in natural language generation, image synthesis, and automated content creation.
* Leveraged GCP AI tools to develop and deploy solutions for NLP, computer vision, and predictive analytics, improving decision-making capabilities across business operations.
* Define and track metrics such as retrieval accuracy, generation quality, relevance, and coherence of RAG responses.
* Performed fine-tuning of LLMs using domain-specific datasets, optimizing hyperparameters and training techniques to enhance model performance
* Integrated various GCP AI services, such as AutoML, Natural Language API, and Vision API, into applications to enhance functionality with pre-trained models for image recognition, text analysis, and language understanding.
* Specialized in developing proof of concept, rapid prototypes using latest AI trends
* Monitored and tuned AI models in GCP AI environments to ensure performance optimization and alignment with business goals, ensuring reliable and efficient AI-powered services.
* Developed prototypes for AI solutions, quickly testing and validating concepts for new products in the Generative AI space, using Python for rapid prototyping and model training.
* Implemented POC related to integration of LLMs and knowledge graphs to improve user search experience and get more accurate results.
* Led the development of POC solutions for personalized content generation using deep learning, evaluating potential business impact and guiding decision-making for future product development.
* Developed and tested NIM models for interactive AI systems, integrating generative capabilities with user input to enable dynamic and responsive AI-driven solutions across industries.
* Executed complex POCs to develop the next gen Conversational AI products with Gen AI features.

**Cognizant (Orlando, FL) April 2018 to October 2021**

**Role: ML Engineer**

* Developed in house No Code Auto ML framework using java, python as per the requirements which automates the entire project life cycle from EDA to deployment.
* Implemented mechanism for generating synthetic data which will be used for training regression models.
* Contributed in converting existing legacy Monolithic application to Micro Services and making it cloud ready. As part of the project, we implemented Java API automation framework and also contributed in API automation.
* Collaborated with business teams to understand data and problem statements and help them to identify, analyze patterns using standard ML approaches.
* Developed and optimized software solutions using Python, Java, and JavaScript, ensuring high performance and scalability across various systems and applications.
* Leveraged advanced analytics techniques to drive data-driven decision and solve problems
* Select appropriate time series models like ARIMA, SARIMA, Prophet as per the requirement and forecast
* Worked on Analyzing, Gain Insights and Trends from data getting generated from application using Athena, EDA techniques. Data insights will be mainly used to make appropriate decisions and improve the existing functionality.
* Conducted retrospectives and process improvements using Agile and DevOps methodologies, resulting in improved development workflows and reduced time-to-market.
* Employed generative AI techniques, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to augment and generate data for machine learning models.
* Perform hyper parameter tuning to optimize ML model performance.
* Developed rapid prototypes and implemented innovative vision based POC spanning object detections and segmentation using python and computer vision techniques.
* Improved model efficiency and accuracy by evaluating developed models.
* Worked on analyzing and building models for the data coming from 3k IOT sensors and contributed for the better performance of the system
* Worked collaboratively with cross-functional teams, including product managers, designers, and developers, to design and develop software applications using Agile and DevOps methodologies.
* Contributed for OCR system to convert the detected text regions into machine-readable text with different languages, handwriting.
* Enhance the performance of OCR systems through advanced techniques and data augmentation to prevent misinterpretation and misalignment issues.
* Use AWS SageMaker Data Wrangler to explore data, identify patterns, and perform EDA using python.
* Contributed in building auto KYC verification system which includes Id proof classification, extraction of text from the proof submitted using various computer vision techniques.
* Work with data lake team to facilitate smooth data flow for model training and maintenance.

**Visam Technologies (Hyderabad, India) June 2016 to March2018**

**Role: ML Engineer**

* Prepared good quality images by using various image enhancement techniques which will be used for model training.
* Developed python scripts to automate tasks and streamline development process.
* Performed data manipulation, data preparation, normalization, predictive modelling.
* Extracted text from images by using OCR techniques and extract features by SIFT algorithm.
* Worked on identifying attributes color variety, color saturation, primary and secondary color from images.
* Developed and implemented predictive models using various ML algorithms.
* Analyze and identify patterns on data sets by applying historical models.
* Worked on data cleaning for the time series data which will be used for forecasting.
* Involved in support of developed ML services in production and resolved production issues by reproducing and determining the root cause analysis.
* Employed DevOps practices, such as continuous integration and continuous delivery, to improve development workflows and reduce time-to-market.
* Implemented LoRA techniques to fine-tune large pre-trained language models, significantly reducing computational costs while maintaining high model performance.
* Worked on the data cleaning part to bring it in proper structure and store it in S3 bucket which will be used for model training.
* Conducted continuous monitoring and tuning of AI models deployed on AWS Bedrock, ensuring optimal performance and minimal latency for real-time use cases.
* Optimized AI/ML workflows using QLoRA, enabling efficient model adaptation with quantization for faster inference and reduced memory footprint
* Written python scripts for image augmentation which will be mainly used to increase the dataset size
* Worked on building various classifiers using VGG as per the given requirement and responsible for improving accuracy and model retraining.
* Conducted ad-hoc data analysis with precision delivering data insights to meet client requirements
* Performed bootstrapping and cross validation for model selection and evaluation.
* Collaborated with various teams to understand data and ensured its effective use in model building.
* Continuously monitor the ML model performance, addressing any issues or anomalies that arise.
* Use SageMaker Experiments to track and compare different training runs, hyper parameter configurations.
* Conducted comprehensive EDA to uncover patterns, trends, and insights within datasets, utilizing statistical methods and visualization tools to guide feature selection, data preprocessing, and model-building decisions