Sindhuri Siddireddy

sindhuri.siddireddy@gmail.com | 203-824-1798 | Pleasanton,CA
https://www.linkedin.com/in/sindhurisiddireddy/

EXPERIENCE

Kaiser Permanente, *Sr ML Engineer* *03/2021 - Present*

- Created AI assistants leveraging Generative AI (GenAI) using Large Language Models (LLMs) to deliver personalized and context-aware responses.
- Leveraged Retrieval-Augmented Generation (RAG) techniques and LangChain to integrate multiple knowledge sources, optimizing the efficiency of information retrieval and enhancing response quality.
- Developed and implemented evaluation pipelines to assess the effectiveness of RAG workflows, focusing on the accuracy, relevance, and speed of information retrieval and response generation.
- Used Hugging Face Transformers to connect models and FAISS and Milvus to quickly and accurately retrieve information in RAG systems.
- Utilized Vector Databases to store and retrieve embeddings efficiently, facilitating quick and accurate similarity searches for improved user query handling.
- Developed a multi-agent system using LangGraph with Agentic AI, to collaborate for complex task automation, improving decision-making and optimizing workflows.
- Designed and implemented prompt engineering strategies such as zero-shot, few-shot and instruction-based prompts, while iteratively refining them to ensure consistent and high-quality outcomes
- Fine-tuned LLMs using domain-specific datasets to enhance AI assistance accuracy, relevance and performance
- Applied model evaluation protocols, including A/B testing , user feedback integration and BLEU, ROUGE, perplexity Score to continuously refine and improve the AI Assistance performance
- Developed NLP workflows using Hugging Face Transformers, NLTK for sentiment analysis and summarization.
- Developed and deployed adaptive machine learning models using PyTorch, implementing real-time classification systems and Automated training, and ensuring scalability in evaluation workflows.
- Engineered high-performance, large-scale ML systems with hands-on experience in PyTorch and TensorFlow, contributing to the design and optimization of neural networks and AI research system
- Evaluated model performance using precision, recall, F1-score, and mean squared error, and applying cross-validation for robustness and reliability.
- Deployed AI assistants on AWS Bedrock, leveraging ECR and EKS and SageMaker for scalable model management
- Monitored model drift, latency, and performance using Datadog, Prometheus, and Grafana for real-time observability and proactive issue detection.
- Implemented CloudWatch-based logging and monitoring for efficient ML workflow troubleshooting.

Whitebox learning *Sr Software Engineer* *01/2018 - 03/2021*

- Designed and implemented end-to-end MLOps pipelines for automating model training, validation, and deployment using Scikit-learn, and MLflow for seamless monitoring.
- Utilized distributed training frameworks to accelerate the training of large-scale ML models, ensuring efficiency.
- Implemented MLOps practices to automate the deployment, monitoring, and management of ML models, ensuring CI/CD , model versioning and performance tracking in production environment
- Designed feature engineering pipelines using Apache Airflow and Databricks to automate data preprocessing for model training and inference.
- Deployed machine learning models in production environments using Docker, Kubernetes, and TorchServe, ensuring scalable and efficient model serving.
- Built intuitive frontend interfaces using React, Redux, and Tailwind CSS to interact with model training, predictions, and analytics.
- Integrated RESTful APIs built with Node.js and Express.js to facilitate seamless communication between ML models and web applications.
- Designed and implemented an optimized Redux store architecture, and Integrated Redux Thunk middleware to enhance the application's ability to handle asynchronous operations within the Redux store and manage complex data flows.
- Designed MongoDB schemas and optimized database interactions for efficient storage and retrieval of ML model metadata, logs, and predictions.
- Worked closely with data scientists, and backend teams to create full-stack solutions that integrate machine learning models with user-friendly web applications.

Cisco Systems Inc. *Full stack Developer* *12/ 2016- 08/2017*

- Developed and integrated responsive React components within the MERN stack, emphasizing custom UI elements to enhance user experience.
- Engineered robust RESTful APIs using Node.js and Express.js, ensuring efficient communication and seamless data flow between the frontend and backend layers.
- Designed efficient MongoDB schemas and implemented CRUD operations for optimized data handling.
- Established secure user authentication with JSON Web Tokens (JWT) for controlled access to application resources.
- Conducted comprehensive testing of MERN stack applications, utilizing Jest and Enzyme for thorough unit testing.
- Collaborated closely with UI/UX designers to implement visually appealing interfaces, utilizing Material-UI and Bootstrap for modern styling and a seamless user journey.
- Introduced and implemented custom UI components, enhancing both the aesthetic and functional aspects of the frontend.
- Contributed to the design and development of reusable React components, fostering a modular and scalable codebase.
- Executed optimization strategies for MongoDB queries, including indexing and aggregation techniques, resulting in significant improvements in data retrieval performance.

Intuit Inc. *Front-End Developer* *03/ 2013 - 09/ 2015*

- Built an AngularJS Single Page Application (SPA) with two-way data binding for seamless synchronization.
- Implemented state-based navigation with UI-Router and encapsulated views with controllers for reusability.
- Utilized AngularJS dependency injection for services and created custom directives for reusable components.
- Designed custom directives with scope isolation, transclusion, and templating for efficient UI development.
- Optimized asynchronous data fetching using Promises ($http, $q), improving frontend rendering speed.
- Used directives for DOM manipulation, filters for formatting, and CSS techniques like sprites and media queries for performance optimization.
- Implemented $scope methods for two-way data binding, ensuring dynamic and interactive user interfaces.

PayPal, Inc. *Front-End Developer* *08/ 2012 - 02/ 2013*

- Designed and developed dynamic web interfaces using HTML, CSS, JavaScript, and jQuery.
- Integrated jqGrid with JSP to enhance admin pages with list and detail popups.
- Used jQuery UI components like Accordion, Autocomplete, Dialog, Menu, Tabs, and Tooltip for improved functionality.
- Built W3C-compliant intranet layouts with CSS for consistency and accessibility.
- Optimized GUI performance with client-side JavaScript and JSP form validations to reduce load times.
- Developed web pages using JavaScript, HTML, and CSS while writing SQL queries for database interactions.
- Integrated frontend views with backend services using AJAX and jQuery for seamless updates.
- Created PHP-based UI components, including jQuery Data Grid and validation controls, for interactive experiences.
- Provided technical support to customers and partners, ensuring timely issue resolution and professional updates.
- Assisted in troubleshooting and resolving customer issues, including on-site support when needed.
- Supported User Acceptance Testing (UAT) during patch releases to ensure software quality.
- Authored and published knowledge base articles to document solutions and best practices.

EDUCATION

Fairfield University *MS in Computer Science*

SKILLS

Gen AI:	LLM, RAG, Vector DB, Langchain, Agentic AI
ML Frameworks/Libraries:	PyTorch, TensorFlow, Scikit-learn, MLOps, Hugging Face
Languages:	Python, JavaScript
Databases:	PostgreSQL, MongoDB, MySQL
Cloud Platforms:	AWS, EKS, S3, SageMaker, Bedrock
Full Stack :	HTML, CSS, JavaScript, React, Angular, Node.js, Redux
Build and CI/CD:	Docker, Kubernetes, Jenkins, GitHub, FastAPI