Narasimulu Alli

Frisco, Dallas, TX 75033. · narasi.mymail@gmail.com · 469-514-1494 ·
https://www.linkedin.com/in/narasimulu-alli/

Education

Osmania University Hyderabad, India
M.Sc IS Master of Science in Information Systems 2001 - 2003

Osmania University Hyderabad, India
BCA Bachelor of Computer Applications 1998 - 2001

Experience

Lucid Motors Dublin, CA
Senior AI Engineer Sep 2024 - Present

- Developed chatbots which use Generative AI (GenAI) and Large Language Models (LLMs) to deliver personalized and context-aware responses, enhancing user engagement and satisfaction.
- Applied NLP techniques using transformer-based models (e.g., BERT, GPT) for natural language understanding and generation, enhancing the chatbot's ability to process and generate contextually accurate responses based on user input and domain-specific language.
- Developed and implemented custom web scraping pipelines using Python libraries (e.g., Scrapy, BeautifulSoup) to gather domain-specific data from websites and APIs, ensuring the chatbot model is continuously fed with high-quality, relevant data to enhance its understanding and response capabilities
- Implemented a Retrieval-Augmented Generation (RAG) framework to improve the chatbot's re- sponse generation by integrating external knowledge sources, thereby enhancing the richness of the interaction.
- Evaluated the RAG framework by assessing retrieval accuracy and response quality, leveraging precision-recall metrics for retrieval and BLEU/ROUGE scores to measure the relevance quality of generated responses.
- Utilized Vector Databases to store and retrieve embeddings efficiently, facilitating quick and accurate similarity searches for improved user query handling.
- Utilized LangChain for advanced conversational capabilities and workflow automation, enhancing the chatbot's ability to manage complex dialogues and tasks.
- Fine-tuned LLMs using domain-specific datasets to optimize the chatbot's performance, ensuring high accuracy and relevance in conversations.
- Integrated Hugging Face Transformers to leverage pre-trained LLMs, streamlining fine-tuning and deployment processes for domain-specific applications, enhancing response quality and model adaptability
- Implemented Agentic AI to enable autonomous decision-making and dynamic task management, enhancing the chatbot's adaptability in real-time interaction.
- Deployed the chatbot on AWS using Bedrock for scalable model training and inference, ensuring reliable and efficient performance under varying load conditions.
- Developed and deployed scalable RESTful APIs with FastAPI, enabling real-time communication between the chatbot and external systems for data retrieval and model inference.
- Designed and managed scalable infrastructure solutions using Amazon EKS for Kubernetes-based container orchestration, ensuring high availability and efficient resource utilization and Git for CI/CD pipeline management.
- Implemented LLMOps pipelines using Docker and custom orchestration scripts, streamlining versioning, deployment, and monitoring workflows specifically tailored for large language models (LLMs), ensuring efficient and reliable production performance.
- Coordinated cross-functional efforts to integrate domain-specific knowledge into Generative AI models, ensuring the system met both technical and business requirements.

CitiCorp Services India Pvt Ltd Pune, India
ML Engineer 2022 - Aug 2024

- Developed and deployed real-time fraud detection models using Python and PyTorch, implementing algorithms for predictive analytics to identify suspicious transactions and enhance security.

- Designed neural networks in PyTorch, utilizing its dynamic computation capabilities to process real-time data efficientl
- Built and maintained end-to-end machine learning pipelines, integrating tools like pandas and scikit-learn for data preprocessing, feature engineering, and model training.
- Implemented MLOps practices using Docker and AWS SageMaker to streamline model deployment, version control, and monitoring in production environments.
- Automated workflows for model training, tuning, and deployment on AWS SageMaker, ensuring scalability and reducing operational overhead.
- Designed and integrated CI/CD pipelines to automate testing, deployment, and updates for machine learning models, ensuring continuous improvement in production systems.
- Evaluated models using ROC-AUC, precision-recall curves, and confusion matrices to ensure robustness in imbalanced datasets.
- Tested and validated models with scikit-learn, ensuring high accuracy and reliability by employing robust evaluation metrics.
- Leveraged monitoring tools on AWS to track performance metrics and ensure system stability with minimal downtime.

CitiCorp Services India Pvt Ltd Pune, India
Senior Software Engineer 2015 - 2022

- Develop, test, and maintain high-quality, scalable software solutions by writing clean and efficient code that aligns with business requirements and technical specifications.
- Work closely with product managers, designers, and fellow engineers to gather requirements, design features, and deliver robust software solutions that meet user needs.
- Identify, troubleshoot, and resolve complex software defects and performance issues to ensure optimal application stability and functionality.
- Participate in thorough code reviews, offering constructive feedback to maintain high coding standards, ensure best practices, and promote knowledge sharing within the team.
- Implement and maintain CI/CD pipelines to automate testing and deployment processes, ensuring faster delivery cycles and minimizing deployment errors.
- Continuously research and evaluate new tools, technologies, and industry trends to incorporate innovative solutions and improve existing systems and processes.

Virtusa Consultancy Services (Polaris) Hyderabad, India / Belfast, UK
Senior Project Lead 2006 - 2015

- Designed and executed infrastructure migration projects, transitioning systems from Solaris/AIX to Linux, ensuring minimal downtime and seamless functionality.
- Developed and maintained implementation plans for project releases, coordinating with cross-functional teams for smooth weekend deployments.
- Conducted disaster recovery (DR) testing in pre-production environments, ensuring system resilience and business continuity.
- Performed infrastructure cost estimation (L0/L1) for new projects, optimizing resource allocation and budget planning.
- Configured and managed Autosys jobs for automated scheduling and monitoring of critical workflows, improving operational efficiency.
- Performed infrastructure cost estimation (L0/L1) for new projects, optimizing resource allocation and budget planning.

Skills

Programming Languages:	Java,Python,SQL
ML Frameworks/Libraries:	Pytorch, Huggingface, Pytorch, MLFlow
Natural Language Processing (NLP:	Huggingface, Pytorch, NLP, Open AI/LLAMA/Anthropic
Frameworks & APIs:	Microservices, REST API, Spring Boot,FastApi
DevOps & CI/CD:	Bitbucket, Jenkins CI/CD pipelines, GitHub, TeamCity, Docker
Other Tools and Technologies:	LLM, Gen AI, RAG, Vector DB, Langchain, Pydantic, Pytest