AI Engineer (2 - 4 Years Experience)

Company Description

Idea Elan LLC is a product based company that provides comprehensive software solutions for research facilities in Universities   and Institutions worldwide.

Please visit www.IdeaElan.com for more information.

Job Description: We are seeking a skilled AI Engineer with 2-4 years of experience to design, develop, and deploy AI/ML solutions, with a strong emphasis on Generative AI, NLP, Retrieval-Augmented Generation (RAG), and Time Series Forecasting. The ideal candidate will have hands-on experience with LangGraph and other Generative AI frameworks to build cutting-edge AI applications.

Key Responsibilities:

1. Develop and deploy AI/ML models with a focus on NLP, Generative AI, RAG, and Time Series Forecasting workflows.
2. Engineer and refine prompts to optimize performance for large language models (LLMs) and generative AI applications.
3. Implement and maintain data preprocessing pipelines, including data cleansing, feature engineering, embedding generation, and transformation for time series data.
4. Develop end-to-end solutions integrating LLMs with VectorDB (e.g., Weaviate, Pinecone, FAISS) for document retrieval, semantic search, and contextual query answering.
5. Utilize LangGraph and other Generative AI frameworks to build structured workflows for LLM-based applications.
6. Train and fine-tune machine learning models, including LLMs, to optimize performance for various use cases.
7. Analyze and interpret complex datasets, including time series data, ensuring scalable and efficient AI model deployment.
8. Collaborate with cross-functional teams to integrate AI/ML models into existing systems and workflows.
9. Ensure models are robust, scalable, and adhere to best practices in ethical AI development.
10. Conduct testing, performance benchmarking, and iterative refinements of AI models and pipelines.
11. Stay updated with the latest advancements in AI/ML, NLP, and cloud technologies and recommend integration strategies for emerging tools and techniques.
12. Create technical documentation and presentations to effectively communicate AI concepts to stakeholders.

Key Skills and Qualifications:

Required Skills:

1. **AI/ML Expertise:** Proven experience in developing, fine-tuning, and deploying AI/ML models, focusing on NLP, Generative AI, and Time Series Forecasting.
2. **Prompt Engineering:** Proficiency in designing, testing, and optimizing prompts for LLMs.
3. **Data Analysis:** Strong ability to process and interpret large and complex datasets, including time series data, for AI model training and validation.
4. **Programming:** Proficiency in Python and experience with AI/ML frameworks like TensorFlow, PyTorch, and scikit-learn.
5. **NLP Techniques:** Understanding of tokenization, embeddings, transformer-based models (e.g., BERT, GPT, LLaMA), and RAG workflows.
6. **Vector Databases:** Experience with Weaviate, Pinecone, FAISS, or similar VectorDBs for document retrieval and storage.
7. **LangGraph & GenAI Frameworks:** Hands-on experience with LangGraph, LangChain, HuggingFace Transformers, OpenAI API, and other Generative AI tools.
8. **Cloud Deployment:** Familiarity with deploying AI solutions on AWS, GCP, or Azure.
9. **Time Series Forecasting:** Experience with ARIMA, LSTM, Prophet, and other forecasting techniques.
10. **Model Training & Fine-Tuning:** Ability to train and fine-tune machine learning models, including large language models (LLMs), to enhance performance and accuracy.
11. **Problem-Solving & Collaboration:** Strong analytical, problem-solving, and teamwork skills to work effectively in a cross-functional environment.

Preferred Skills:

1. Experience in implementing RAG workflows and integrating generative AI with retrieval-based systems.
2. Familiarity with ethical AI principles and compliance frameworks.
3. Knowledge of additional programming languages such as JavaScript or SQL.
4. Proficiency in MLOps practices for automating AI/ML pipelines and lifecycle management.
5. Experience in developing AI-driven applications for real-world industry use cases.