**Generative AI Engineer**

**Position Summary:**

We are seeking a highly skilled and innovative Generative AI Engineer to join our dynamic team. This role is ideal for someone passionate about pushing the boundaries of artificial intelligence, particularly in conversational AI and knowledge-based systems. The candidate will focus on developing, designing, and maintaining cutting-edge AI-based systems, ensuring smooth and engaging user experiences. Additionally, the Generative AI Engineer will participate in a wide variety of Natural Language Processing activities, including refining and optimizing prompts to improve the outcome of Large Language Models (LLMs), and code and design review.

**Responsibilities:**

* Working across client teams to develop and architect Generative AI solutions using ML and GenAI
* Evaluating and selecting appropriate AI tools and machine learning models for tasks, as well as building and training working versions of those models using Python and other open-source technologies
* Working with leadership and stakeholders to identify AI opportunities and promote strategy.
* Building and prioritizing backlog for future machine-learning enabled features to support client business processes.
* You'll design and build generative models, selecting the most suitable architecture (e.g., GANs, VAEs) based on the desired output (text, images, code). This involves writing code using Python libraries like TensorFlow or PyTorch.
* Once your model is built, you'll train it on the prepared data, fine-tuning hyperparameters to achieve optimal performance. You'll then evaluate the model's outputs to assess its effectiveness and identify areas for improvement.
* You'll collaborate with other engineers to integrate your generative AI solution into existing systems or develop new applications. This might involve deploying the model on cloud platforms for scalability.
* The field of generative AI is rapidly evolving. Staying abreast of the latest research, advancements, and ethical considerations in AI development is an ongoing process.

**Requirements:**

* 3+ years of experience programming in Python or R.
* Knowledge of Python libraries like Pandas, Scikit-Learn, Numpy, NLTK is required
* 3+ years of experience with Natural Language Processing (NLP) and Large Language Models (LLM) 3+ years of experience building and maintaining scalable API solutions
* Experience working with RAG technologies and LLM frameworks (Langchain, Claude and LLamaIndex), LLM model registries (Hugging Face), LLM APIs, embedding models, and vector databases (FAISS , Milvus , OpenSearch, Pinecone etc.)
* Experience working with cloud platforms (e.g., AWS, Azure, Google Cloud) and containerization technologies (e.g., Docker, Kubernetes).
* Utilize tools such as Docker, Kubernetes, and Git to build and manage AI pipelines
* Experience driving DevOps and MLOps practices, covering continuous integration, deployment, and monitoring of AI
* 3+ years of experience in data wrangling/cleansing, statistical modeling, and programming
* Familiarity with machine learning pipelines (Azure ML)
* Deep understanding of generative AI technologies and their underlying mechanisms; excellent problem-solving skills and creativity in designing prompts for complex queries
* Exceptional collaboration and communication skills, with the ability to work effectively in a team and articulate ideas to both technical and non-technical stakeholders; agility and adaptability in a fast-paced, evolving technology landscape