Senior AI Research Engineer

**Job Overview:** We are seeking a **Senior AI Research Engineer** with expertise in **transformers, PyTorch, NLP, generative AI**, and **large language models (LLMs)**. You will work in a dynamic environment focused on **Azure OpenAI**, **Assistant API**, **Chat Completions**, and **LLM integration (Google/OpenAI)**. Familiarity with **Python** and proficiency in multiple programming languages will be a significant advantage. This role is key in shaping our AI-powered solutions, from chatbot automation to advanced web scraping techniques.

**Key Responsibilities:**

* Advanced Model Architecture: Develop and optimize transformer-based models for NLP tasks, ensuring scalability and efficiency in the cloud environment (Azure/AWS).
* Generative AI Development: Build and refine generative models, particularly LLMs, to generate coherent and contextually appropriate text for applications like customer care and chatbot automation.
* AI Integration: Work on integrating Azure OpenAI, Assistant API, and other LLM solutions into practical AI-driven products, such as chatbots and virtual assistants.
* Research and Innovation: Lead AI research initiatives focusing on cutting-edge transformer and NLP methodologies, exploring new applications and improving model performance.
* Technical Leadership: Provide leadership in PyTorch, transformers, and LLMs, guiding the development team through complex training and deployment challenges.
* Collaboration & Mentorship: Collaborate with cross-functional teams to incorporate AI technologies into diverse products and mentor junior engineers and researchers.

**Qualifications:**

*Education:*

* A PhD in Computer Science, Artificial Intelligence, Machine Learning, or a related field (preferred but not required). Equivalent experience with strong practical exposure is highly valued.

*Experience:*

* Proven hands-on experience in transformer architectures, LLMs, and NLP.
* Demonstrated success in building and deploying large-scale generative AI models using PyTorch.
* LLM integration experience with Google and OpenAI.
* Familiarity with Assistant APIs and Chat Completions technologies.
* Experience in multi-cloud environments, including Azure and AWS.

*Technical Skills:*

* Deep expertise in NLP and generative AI, particularly with transformer-based architectures.
* Proficiency in PyTorch, understanding its mechanisms and optimization techniques.
* Hands-on knowledge of Azure OpenAI and Assistant APIs for developing AI-powered solutions.
* Advanced techniques in model fine-tuning, regularization, and LLM interpretability.
* Programming expertise in Python and at least one additional language.

**Preferred Qualifications:**

* Published research or contributions to open-source AI projects.
* Experience in cloud-based AI model deployment on platforms like Azure, AWS, or Google Cloud.
* Familiarity with AI-driven automation solutions (e.g., chatbots, customer care automation).