**Job Posting: Retrieval-Augmented Generation (RAG) Application Developer**

| **Position Title** | **RAG Application Developer / LLM Engineer** |
| --- | --- |
| **Department** | Artificial Intelligence / Data Science |
| **Location** | [Remote / Hybrid / On-site – City, Country] |
| **Employment Type** | Full-time |
| **Experience Level** | 2–6 years (flexible based on expertise) |
| **Reports To** | Lead – AI Systems / Director, ML Engineering |
| **Application Deadline** | [Insert Date] |
| **Contact Email** | [Insert HR / Hiring Contact] |

**About the Role**

We are looking for a **RAG Application Developer** to design, develop, and optimize **Retrieval-Augmented Generation** systems that integrate large language models (LLMs) with real-time data sources. The candidate will work on building scalable, production-grade AI pipelines that enhance knowledge discovery and automate reasoning across structured and unstructured data.

**Key Responsibilities**

* Design and implement **RAG pipelines** combining retrieval and generation modules.
* Develop and optimize **document retrieval** components using vector databases (FAISS, Pinecone, Chroma, Milvus, etc.).
* Build **custom embedding generation workflows** for text and multimodal data.
* Integrate RAG frameworks like **LangChain, LlamaIndex, Haystack**, or in-house orchestration layers.
* Fine-tune system performance through **retrieval quality metrics and evaluation datasets**.
* Deploy, monitor, and maintain RAG systems on **cloud platforms (AWS, GCP, Azure)**.
* Collaborate with backend teams to expose model capabilities via **APIs and microservices**.
* Contribute to internal tools and libraries for **LLM pipeline automation** and **knowledge retrieval**.

**Required Skills and Qualifications**

* Bachelor’s/Master’s in Computer Science, Data Science, or related field.
* Proficiency in **Python**, **LLM APIs**, and **retrieval frameworks**.
* Experience with **LangChain**, **LlamaIndex**, or **Haystack** for RAG implementations.
* Hands-on with **vector databases** like Pinecone, Weaviate, or Chroma.
* Understanding of **transformer architectures** and **embedding models** (OpenAI, Sentence Transformers, Cohere).
* Familiarity with **prompt engineering**, **context chunking**, and **retrieval evaluation**.
* Knowledge of **containerization (Docker, Kubernetes)** and **RESTful API integration**.
* Strong analytical and debugging skills with a focus on production readiness.

**Preferred Qualifications**

* Experience building **enterprise-grade or domain-specific RAG systems**.
* Familiarity with **multimodal RAG** (text + image + structured data).
* Exposure to **fine-tuning** or **parameter-efficient tuning** (LoRA, PEFT).
* Prior contributions to **open-source LLM/RAG projects**.
* Experience with **model monitoring and observability** in production.

**What We Offer**

* Work on **state-of-the-art AI and RAG systems** with real-world impact.
* Collaborative, innovative, and research-friendly environment.
* Competitive compensation and professional development opportunities.
* Flexible hybrid/remote work structure.

**Application Process**

Please submit the following to [Insert Contact Email or Application Link]:

1. Resume/CV highlighting RAG and LLM-related experience
2. Links to GitHub, portfolio, or prior AI projects
3. A short note (2–3 lines) describing your interest in RAG-based systems

Would you like me to make a **version tailored for government/research institute use** (e.g., ICAR, NIC, or academic research lab format with project duration, funding source, and grade pay)?  
That version would include a second table for **Project Details** and **Tenure**, common in official postings.