Here’s a clean, ready-to-use **job posting document content** for a **RAG (Retrieval-Augmented Generation) Application Developer** role. You can use it as-is for internal posting, LinkedIn, or official documentation.

**Job Title:**

RAG Application Developer / LLM Engineer (Retrieval-Augmented Generation)

**Location:**

[Specify: Remote / Hybrid / On-site – City, Country]

**About the Role:**

We are seeking a skilled **RAG Application Developer** to design, build, and optimize applications powered by **Retrieval-Augmented Generation (RAG)** techniques. The ideal candidate will have strong experience in integrating **Large Language Models (LLMs)** with structured and unstructured data sources to enable intelligent, context-aware information retrieval and response generation.

You will work closely with AI researchers, data engineers, and backend teams to deliver production-grade AI systems that enhance knowledge discovery, automate reasoning, and provide real-time insights.

**Key Responsibilities:**

* Design and implement **RAG pipelines** that connect LLMs with document databases, APIs, and enterprise data sources.
* Develop **retrieval and ranking components** using vector databases (e.g., FAISS, Pinecone, Chroma, Milvus).
* Build **custom embeddings** pipelines for text, images, or multimodal data using transformer-based models.
* Integrate **context retrieval** mechanisms into LLM workflows using frameworks like **LangChain, LlamaIndex**, or custom orchestration layers.
* Fine-tune and evaluate model performance using metrics such as retrieval accuracy, relevance, and response coherence.
* Deploy, monitor, and optimize RAG systems for scalability, latency, and cost efficiency.
* Collaborate with backend teams to expose RAG capabilities via REST or GraphQL APIs.
* Contribute to internal tools and libraries for RAG experimentation, data curation, and knowledge management.

**Required Skills and Qualifications:**

* Bachelor’s or Master’s in Computer Science, Data Science, AI, or related fields.
* Strong programming skills in **Python** and familiarity with frameworks like **LangChain**, **LlamaIndex**, or **Haystack**.
* Solid understanding of **transformers**, **vector search**, and **embedding models** (e.g., OpenAI, Cohere, Sentence Transformers).
* Experience with **LLMs (GPT, Claude, LLaMA, Gemini, etc.)** and prompt engineering.
* Hands-on experience with **vector databases** such as Pinecone, Weaviate, Milvus, or Chroma.
* Familiarity with **document loaders, chunking strategies**, and **retrieval evaluation techniques**.
* Knowledge of **cloud platforms** (AWS, GCP, Azure) and containerization (Docker, Kubernetes).
* Strong debugging, analytical, and communication skills.

**Preferred Qualifications:**

* Experience building **enterprise RAG systems** with security and access control.
* Familiarity with **multimodal RAG** (text, image, or structured data).
* Exposure to **fine-tuning, LoRA, or PEFT** techniques for domain-specific adaptation.
* Prior work with **open-source LLMs** and **embedding model customization**.
* Contributions to **open-source RAG or LLM frameworks**.

**What We Offer:**

* Opportunity to work with cutting-edge **LLM and RAG technologies**.
* Collaborative, research-driven environment.
* Competitive compensation and benefits package.
* Flexible work model and career growth opportunities.

**Application Instructions:**

To apply, please submit:

* Your updated resume
* Links to GitHub, publications, or portfolio (if available)
* A short note (2–3 lines) on your experience with RAG or LLM-based systems

Would you like me to tailor this for a **specific organization type** (e.g., startup, government lab, enterprise AI team, or research institute)?  
I can make it more formal, academic, or industry-oriented depending on your context.