**Assignment: Integration of Retrieval Augmented Generation (RAG) with Open-Source LLM and LangChain for Autism Intervention Research**

**Aryan Singh Kushwaha**

**Introduction**

In this assignment, I explore the integration of retrieval augmented generation (RAG) with the open-source LLM Mistral-7B-Instruct-v0.1 and the LangChain library to address a specific NLP task within the context of autism intervention research.

**Methodology**

* Document Preparation:
  + PyPDF2: Used for efficient parsing of PDF-based research documents.
  + Regex and NLTK: Employed for text cleaning and preprocessing (removing citations, links, email IDs).
  + LangChain Text Splitter: Ensured consistent chunk sizes before database upload.
* Vector Database:
  + ChromaDB: Selected for its ease of use and embedded vectorization capabilities (using the default all-MiniLM-L6-v2 model).
* LLM Integration:
  + Transformers and PyTorch: Leveraged to construct a pipeline for Mistral-7B-Instruct-v0.1.
  + LangChain: Crucially, LangChain provided:
    - Seamless integration with HuggingFace models.
    - Construction of LLM chains for task execution.
    - Flexible prompt templating for query-based interactions.
* RAG Implementation:
  1. ChromaDB Query: LangChain-facilitated query used to retrieve top-5 relevant research snippets.
  2. Intermediate Summarization: LLM chain generated concise summaries of each retrieved snippet.
  3. Final Response Generation: Main LLM chain, fed with these summaries, synthesized the final response, informed by the most pertinent research context.

**LangChain's Use,**

Abstraction: LangChain effectively abstracted low-level interactions with LLMs and the vector database, simplifying the development process.

Modularity: Its modular design enabled the construction of a complex RAG system through a clear, chain-based workflow.

Flexibility: LangChain's adaptable prompt templates streamlined the process of querying and prompting different stages of the RAG process.

**Possible areas of improvement:**

* Using LLMs to clean extracted text to suite user needs
* Creating a conversation chains with various memory buffers to allow LLM to remember Autism Intervention Research related content.