## **Automated Assessment with Retrieval-Augmented Generation (RAG)**

**Part 1: What is RAG and Where Does it Come From?**

Welcome everyone! Today, we'll explore Retrieval-Augmented Generation (RAG), a revolutionary technique for enhancing Large Language Models.

LLMs are powerful tools trained on massive datasets to perform tasks like text generation, language translation, and question-answering. However, a significant challenge arises when dealing with factual accuracy. LLMs rely solely on the information they were trained on, leading to consistency and limitations when encountering complex subjects.

RAG tackles this challenge by introducing an "information retrieval" component. RAG can be fine-tuned, and its internal knowledge can be modified efficiently without retraining the entire model.

RAG takes input and retrieves relevant/supporting documents from a source, e.g., Wikipedia. The documents are concatenated in context with the original input prompt and fed to the text generator, which produces the final output. This makes RAG adaptive for situations where facts could evolve. This is very useful as LLMs's parametric knowledge is static. RAG allows language models to bypass retraining, enabling access to the latest information for generating reliable outputs via retrieval-based generation.

The concept of RAG is relatively young, emerging in 2020 from the minds of researchers at Meta AI. They recognised the potential of RAG to improve the reliability and factual grounding of LLM-generated responses significantly.

One key advantage of RAG is its fine-tunability. Unlike traditional LLMs that require extensive retraining for new information, RAG allows for efficient modification of its internal knowledge base. This is achieved by updating external knowledge sources, making RAG highly adaptable to evolving factual landscapes.

Here's a deeper dive into how RAG works:

1. **Input and Retrieval:** The system receives the student's answer as input. RAG then activates its information retrieval component, scouring the designated knowledge base for relevant documents that complement or support the answer.
2. **Contextual Concatenation:** The retrieved documents are strategically combined with the original input prompt. This creates a richer context for the LLM to analyse, ensuring a more comprehensive understanding of the student's response.
3. **Text Generation with Enhanced Knowledge:** This enriched context is then fed to the text generator component, which is essentially Mistral 7b. However, with RAG, Mistral 7b leverages its internal knowledge and the retrieved information to generate a more accurate and factually grounded response when evaluating student answers.

**Part 2: Why Use RAG? Boosting Efficiency and Accuracy**

Now, let's delve into the significance of RAG for our automated assessment system.

* **Enhanced Factual Accuracy:** As discussed, RAG empowers Mistral 7b to access and utilise factual information from external sources. This ensures a higher level of accuracy when evaluating student responses, particularly in subjects like history, physics, or biology. Imagine a biology question about cell structures. Mistral 7b, without RAG, might rely on statistical patterns in its training data, potentially leading to inaccurate assessments. However, with RAG, it can access scientific databases, guaranteeing the response aligns with established scientific facts. This is crucial for ensuring fair and objective evaluation of student understanding.
* **Improved Efficiency and Flexibility:** Constantly retraining a large LLM like Mistral 7b for every new subject or question can be cumbersome and time-consuming. RAG elegantly bypasses this by integrating external knowledge. By adding new relevant knowledge bases, we can adapt the system to a wider range of assessment topics, saving valuable time and resources. Imagine incorporating a new subject like economics into the assessment system. With traditional LLMs, we might need to retrain the entire model. However, with RAG, we simply add a curated database of economic resources, allowing the system to assess student responses in this new domain effectively.
* **Providing Up-to-date and Accurate Responses:** The real world constantly evolves, and knowledge can become outdated. Traditional LLMs, with their static knowledge base, struggle to keep pace. RAG, however, ensures the system's responses are based on the latest information by dynamically accessing external data sources. This is particularly valuable in fields like medicine or technology, where rapid advancements happen.

## 

**Part 3: How RAG Benefits Our Automated Assessment System**

* **Addressing Open-Ended Responses:** As previously mentioned, RAG empowers Mistral 7b to delve deeper into student responses, especially for open-ended questions. By retrieving factual information related to the topic, RAG provides valuable context for evaluating the student's reasoning and understanding. Imagine a question asking students to analyse the social and political factors leading to a historical event. Mistral 7b, without RAG, might struggle to assess the depth and accuracy of the student's explanation. However, with RAG, it can retrieve relevant historical documents and analyse the student's response against that factual backdrop. This allows for a more nuanced evaluation of the student's critical thinking skills.
* **Identifying Factual Errors and Providing Learning Opportunities:** RAG is crucial in identifying potential factual errors in student responses. Imagine a student mistakenly mentions the wrong scientific theory in a biology question. RAG can highlight this discrepancy, allowing the system to flag the answer for further review. However, RAG goes beyond simply identifying errors. It can also offer students access to the retrieved information that corrects the mistake. This transforms the assessment process into a learning opportunity, allowing students to identify and rectify their knowledge gaps.
* **Reduced Bias and Fairer Evaluation:** Traditional LLMs trained on massive datasets can sometimes inherit biases in that data. RAG helps mitigate this by allowing us to curate the external knowledge base, ensuring it represents a diverse and objective range of perspectives. This fosters fairer evaluation for students from all backgrounds.

**Conclusion**

RAG is not just a technical advancement; it's a game-changer for LLM-based systems like ours. By incorporating factual accuracy, efficiency, adaptability, and the ability to handle diverse question types, RAG empowers our automated assessment system to provide a more reliable, effective, and fair evaluation experience for students and teachers. It opens doors for personalised learning opportunities and a more robust assessment environment.