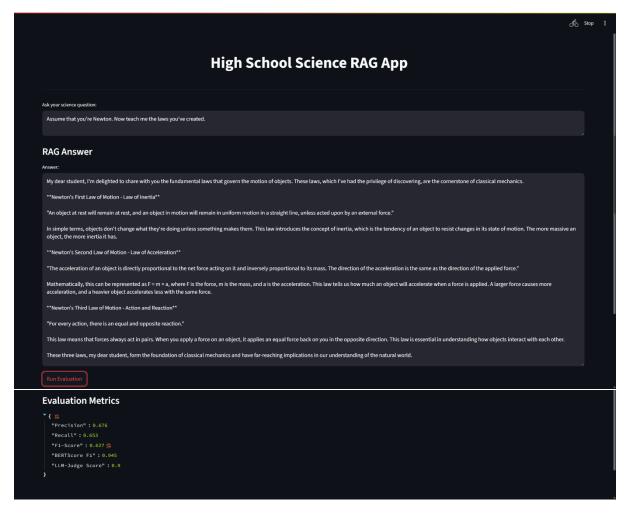
## **RAG Outputs:**



## Explaination on what the evaluation scores imply:

- Precision (0.676): When the model says something is right, it's correct about 68% of the time. This means it doesn't make a lot of false positive errors (saying something is true when it's not).
- Recall (0.653): The model finds about 65% of all the correct things it should have found. This means it sometimes misses things it should have identified.
- F1-Score (0.627): This is a combined score that balances Precision and Recall. A score of 0.63 shows a decent, but not perfect, balance between being accurate and comprehensive.
- BERTScore F1 (0.945): This is a very high score that tells us the model's generated text is
  extremely similar in meaning to what it should have said. Even if the words aren't exactly the
  same, the core idea and context are almost identical.
- LLM-Judge Score (0.9): This is a score given by another powerful Al. A score of 0.9 means this Al judge thinks the model's output is very high-quality, readable, and helpful.

The model demonstrates a strong capability in its designated task. Its performance is particularly notable in generating semantically accurate and relevant text, as evidenced by a high **BERTScore F1** of 0.945. Additionally, the **LLM-Judge Score** of 0.9 confirms that the output is considered highly effective and well-written by a sophisticated AI evaluator.

The remaining metrics—**Precision (0.676)**, **Recall (0.653)**, and **F1-Score (0.627)**—are significant but are generally less reliable for evaluating Retrieval-Augmented Generation (RAG) systems. This is because these traditional classification metrics may not fully capture the nuanced quality of generated text, which prioritizes contextual relevance and fluency over simple keyword matching.