The correct evaluation of the system requires a structured approach across several key areas:

- **Accuracy of Fact-Checking**
  The system's fact-checking results should be compared against a labeled dataset to assess how well it distinguishes between true and false claims. Standard evaluation metrics such as **precision, recall, and F1-score** provide a quantitative measure of its reliability.
- **Effectiveness of Similarity Matching**
  The cosine similarity threshold plays a crucial role in retrieving relevant fact-checked claims. If set too high, the system may miss useful matches; if too low, it may retrieve irrelevant ones. Fine-tuning this parameter ensures an optimal balance between precision and recall.
- **LLM Response Quality**
  When no fact-check is found, the system relies on an LLM to generate a response. Evaluating the factual correctness and usefulness of these responses is essential. Human review remains the most reliable method to determine response quality.
- **Speed & Latency**
  A fact-checking system must be both accurate and efficient. Measuring the response time for retrieving fact-checks and generating LLM-based responses helps identify potential bottlenecks. Optimizing performance ensures a seamless user experience.
- **User Feedback & Error Analysis**
  Real-world testing with users provides valuable insights into the system's strengths and weaknesses. Analyzing incorrect fact-checks and refining the retrieval or LLM response mechanisms ensures continuous improvement.

A combination of automated evaluation and real-world testing is necessary to ensure that the system remains **accurate, efficient, and practical** for users.