Llama Index Evaluation Test Runs

Though there is Inconsistency in the Output and Metric's Values but these are the most consistent results so far.

Test Case 1: Static / Defined Input

Setup:

1. **Queries**: Passed three queries
2. **Prompt**:
   "You are a helpful assistant. Answer the question to the best of your ability. If you don't know the answer, say 'I am not sure about that.'"
3. **Context for Response Generation**:
   a. Context is passed only for certain queries
   b. **No Context** = empty string **""**
   c. **Incorrect Context** = context unrelated to the query
4. **Context for Evaluation**:
   a. Same context passed for both generation and evaluation

Observations:

- For **Correct Context**: All metrics (Correctness, Faithfulness, Relevance) passed
- For **Incorrect Context**: Only **Faithfulness** failed
- For **No Context**: Only **Faithfulness** failed

Metric Behavior Summary:

- **Correctness**:
  - Independent of context
  - Dependent on query and response
  - Highly consistent
- **Relevance**:
  - Independent of context
  - Dependent on query and response
  - Highly consistent
- **Faithfulness**:
  - Dependent on both context and response
  - Consistent in behavior

Test Case 2: No Context or Wrong Context Passed

Inconsistencies have been observed in this test case due to Dynamic Input

Setup:

1. **Input**: Dynamic (user inputs three queries)
2. **Prompt**:
   "You are a helpful assistant. Answer the question to the best of your ability. If you don't know the answer, say 'I am not sure about that.'"
3. **Context for Response Generation**: Not passed
4. **Context for Evaluation**: Passed
   a. Context was only relevant for **Query 2**

Output:

Query 1 – "What is the capital of France?"
Context: Irrelevant

- Correctness: **True** (4/4)
- Relevance: **True** (1/1)
- Faithfulness: **False** (0/1)

Query 2 – "What is the capital of Pakistan?"
Context: Correct

- Correctness: **True** (4/4)
- Relevance: **True** (1/1)
- Faithfulness: **True** (1/1)

Query 3 – "Do we need air to breathe?"
Response: Lengthy and complex
Context: Irrelevant

- Correctness: **False** (3/4)
- Relevance: **False** (0/1)

- Faithfulness: **False** (0/1)

Metric Analysis

Correctness

- Independent of context
- Evaluated using only the query and response
- Inconsistency in scoring observed (e.g., range 0–4 instead of binary)
- May score lower if response is complex or lengthy
- LLM model can influence output

Relevance

- Intended to be context-independent
- Depends on response relevance to the query
- Sometimes returns `0` or `N/A` when no context is passed
- May return false if context is irrelevant, despite being designed to ignore context

Faithfulness

- Depends on both context and response
- Returns false if context is:
    - Missing
    - Empty
    - Irrelevant
- Unclear if LLM's own internal knowledgebase qualifies as valid "context"

Final Summary Table

Metric	Context Dependency	Evaluation Basis	Behavior
Correctness	No	Response only	Slightly inconsistent (due to response length/complexity)
Relevance	No (intended)	Response only	Inconsistent when context is wrong or missing

Faithfulness	Yes	Context + Response	Fails when context is wrong, empty, or not passed