Analytics Vidhya

Correctness

Instructors

Prashant Sahu

Manager - Data Science, Analytics Vidhya

Ravi Theja

Developer Advocate Engineer, LlamaIndex

Evaluation Metrics

- Retriever Evaluation Metrics
  - Hit rate
  - Mean Reciprocal Rank (MRR)
- Response Evaluation Metrics
  - Faithfulness
  - Relevancy
  - **Correctness**
  - Semantic Similarity
  - Adherence Guideline

Correctness

Defines the accuracy of responses to the actual responses

Correctness

Uber Technologies annual/quarterly revenue history and growth rate from 2017 to 2023.

- Uber Technologies revenue for the quarter ending September 30, 2023 was $9.292B, a 11.37% increase year-over-year.

- Uber Technologies revenue for the twelve months ending September 30, 2023 was $35.952B, a 23.77% increase year-over-year.

- Uber Technologies annual revenue for 2022 was $31.877B, a 82.62% increase from 2021.

- Uber Technologies annual revenue for 2021 was $17.455B, a 56.7% increase from 2020.

Correctness

Uber Technologies annual/quarterly revenue history and growth rate from 2017 to 2023.

- Uber Technologies revenue for the quarter ending September 30, 2023 was $9.292B, a 11.37% increase year-over-year.

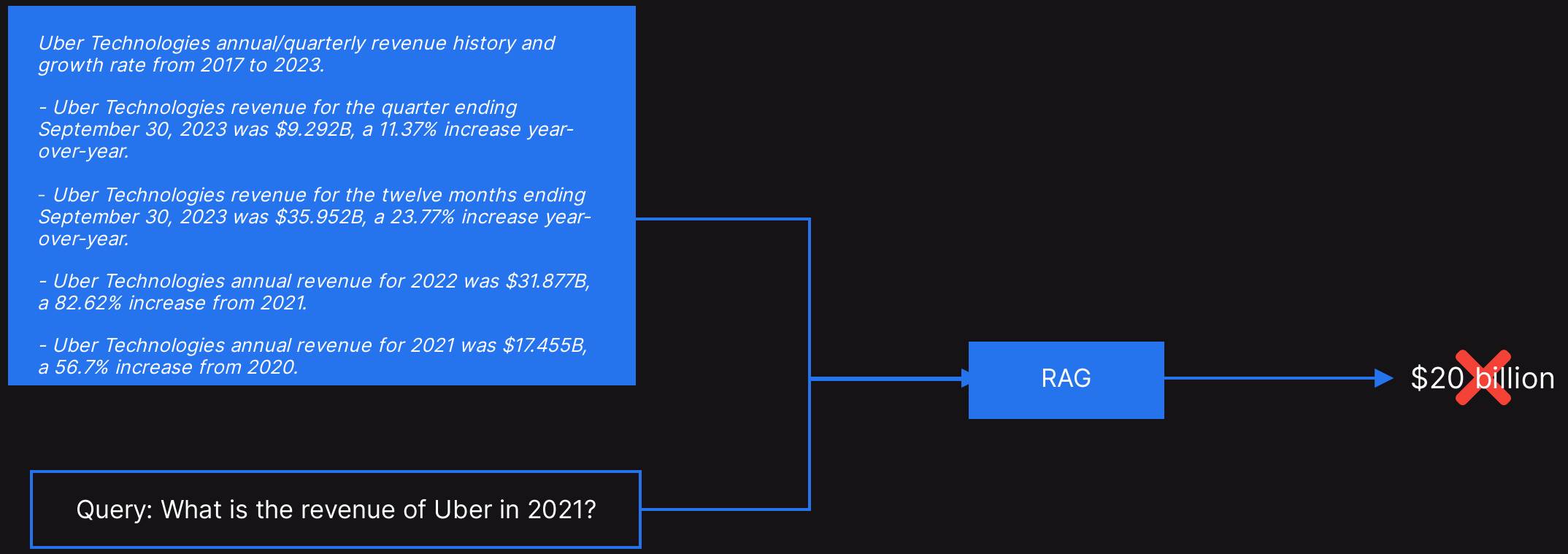
- Uber Technologies revenue for the twelve months ending September 30, 2023 was $35.952B, a 23.77% increase year-over-year.

- Uber Technologies annual revenue for 2022 was $31.877B, a 82.62% increase from 2021.

- Uber Technologies annual revenue for 2021 was $17.455B, a 56.7% increase from 2020.

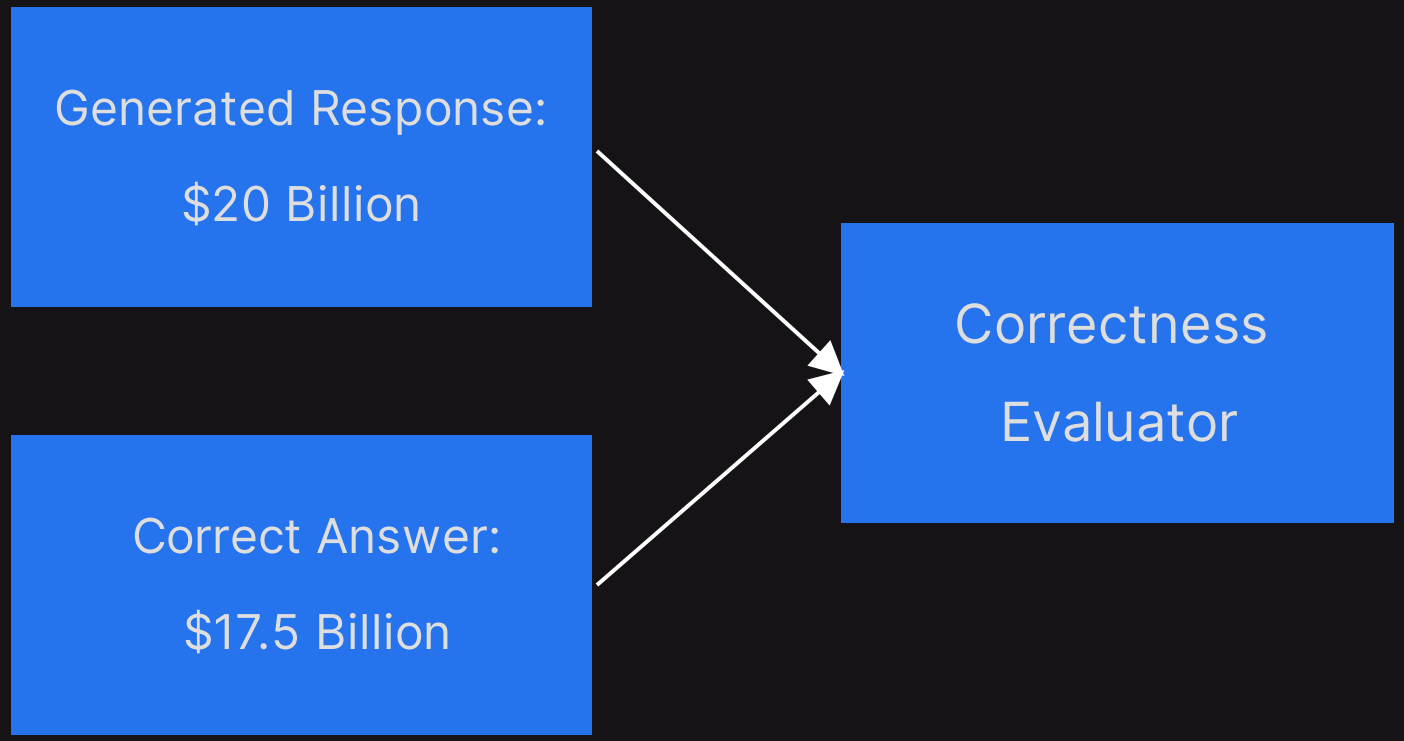
Query: What is the revenue of Uber in 2021?

Correctness



Analytics Vidhya

Correctness Evaluator



- Feedback
- Score between 1 to 5
- Passing (Yes/ No)

```
from llama_index.core.evaluation import CorrectnessEvaluator
```

Calculating Correctness Score

- Answer Correctness measures the accuracy of the generated answer when compared to the ground truth.
- Correctness Score ranges between 0 to 1
- Correctness Score has 2 components:
  - Factual Correctness
  - Semantic Similarity

(A) Calculating Factual Correctness

- Factual correctness quantifies the factual overlap between the generated answer and the ground truth answer.
- Components used to calculate Factual correctness:
  - **TP (True Positive)**: the ground truth and the generated answer.
  - **FP (False Positive)**: the generated answer but not in the ground truth.
  - **FN (False Negative)**: the ground truth but not in the generated answer.

(A) Calculating Factual Correctness: Example

Example Scenario:

Ground truth: Einstein was born in 1879 in Germany.

Answer 1 (High answer correctness):
In 1879, Einstein was born in Germany.

Answer 2 (Low answer correctness):
Einstein was born in Spain in 1879.

- In the Answer 2:
  - TP: [Einstein was born in 1879]
  - FP: [Einstein was born in Spain]
  - FN: [Einstein was born in Germany]
- We then calculate F1 score

(A) Calculating Factual Correctness: F1-Score

Example Scenario:

Ground truth: Einstein was born in 1879 in Germany.

Answer 1 (High answer correctness):
In 1879, Einstein was born in Germany.

Answer 2 (Low answer correctness):
Einstein was born in Spain in 1879.

- In the Answer 2:
  - TP: [Einstein was born in 1879]
  - FP: [Einstein was born in Spain]
  - FN: [Einstein was born in Germany]

$$\text{F1 Score} = \frac{|\text{TP}|}{(|\text{TP}| + 0.5 \times (|\text{FP}| + |\text{FN}|))}$$

(B) Calculating Semantic Similarity

- This evaluation is based on the ground truth and the answer, with values in the range of 0 to 1.
- A higher score signifies a better alignment between the generated answer and the ground truth.
- This evaluation utilizes an Embedding model to calculate the semantic similarity score.

(B) Calculating Semantic Similarity: Calculation

Hint

Ground truth: Albert Einstein's theory of relativity revolutionized our understanding of the universe."

High similarity answer: Einstein's groundbreaking theory of relativity transformed our comprehension of the cosmos.

Low similarity answer: Isaac Newton's laws of motion greatly influenced classical physics.

- **Step 1:** Vectorize the ground truth answer using an embedding model.
- **Step 2:** Vectorize the generated answer using the same embedding model.
- **Step 3:** Compute the cosine similarity between the two vectors.

(B) Calculating Semantic Similarity: Calculation

- **Step 1:** Vectorize the ground truth answer using the specified embedding model.
- **Step 2:** Vectorize the generated answer using the same embedding model.
- **Step 3:** Compute the cosine similarity between the two vectors.

$$Semantic\ Similarity = Cosine\_Similarity(ground\_truth\_vector, generated\_answer\_vector)$$

Calculating Answer Correctness

$$Answer\ Correctness = (W_1 * Factual\ Correctness\ score) + (W_2 * Semantic\ Similarity)$$

$$where\ W_1\ and\ W_2\ are\ weights$$

Factors to improve correctness

- Chunk size
- Embedding model
- Retrieval algorithm
- Response synthesis LLM

Thank You

Analytics Vidhya