

LLM Tutorials

RAG evals

Metrics to evaluate a RAG system.

In this tutorial, we'll demonstrate how to evaluate different aspects of Retrieval-Augmented Generation (RAG) using Evidently.

- ❗ We'll demonstrate a **local open-source workflow**, viewing results as a pandas dataframe and a visual report — ideal for Jupyter or Colab. At the end, we also show how to upload results to the Evidently Platform. If you are in a non-interactive Python environment, choose this option.

We will evaluate both retrieval and generation quality:

Retrieval. Assessing the quality of retrieved contexts, including per-chunk relevance.

Generation. Evaluating the quality of the final response, both with and without ground truth.

By the end of this tutorial, you'll know how to evaluate different aspects of a RAG system, and generate structured reports to track RAG performance.

- ❗ Run a sample notebook: [Jupyter notebook](#) or [open it in Colab](#).

- ❗ To simplify things, we won't create an actual RAG app, but will simulate getting scored outputs. If you want to see an example where we also create a RAG system, check this [video tutorial](#).

1. Installation and Imports

Install Evidently:

```
!pip install evidently[llm]
```





≡ LLM Tutorials > RAG evals

```
from evidently import Dataset
from evidently import DataDefinition
from evidently.descriptors import *

from evidently import Report
from evidently.presets import TextEvals
from evidently.metrics import *
from evidently.tests import *

from evidently.ui.workspace import CloudWorkspace
```

Pass your OpenAI key as an environment variable:

```
import os
os.environ["OPENAI_API_KEY"] = "YOUR_KEY"
```



2. Evaluating Retrieval

Single Context

First, let's test retrieval quality when a single context is retrieved for each query.

Generate a synthetic dataset. We create a simple dataset with questions, retrieved contexts, and generated responses.



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



☰ LLM Tutorials > RAG evals



≡ LLM Tutorials > RAG evals

```
synthetic_data = [  
    ["Why do flowers bloom in spring?",  
     "Plants require extra care during cold months. You should keep them indoors."  
     "because of the rising temperatures"],  
    ["Why do we yawn when we see someone else yawn?",  
     "Yawning is contagious due to social bonding and mirror neurons in our brains"  
     "because it's a glitch in the matrix"],  
    ["How far is Saturn from Earth?",  
     "The distance between Earth and Saturn varies, but on average, Saturn is abou"  
     "about 1.4 billion kilometers"],
```



≡ LLM Tutorials > RAG evals

```
columns = ["Question", "Context", "Response"]
synthetic_df = pd.DataFrame(synthetic_data, columns=columns)
```

To be able to preview a full-width pandas dataset.

```
pd.set_option('display.max_colwidth', None)
```

Evaluate overall context quality. We first assess whether the retrieved context provides sufficient information to answer the question and view results as a pandas dataframe.

```
context_based_evals = Dataset.from_pandas(
    synthetic_df,
    data_definition=DataDefinition(text_columns=["Question", "Context", "Response"],
    descriptors=[ContextQualityLLMEval("Context", question="Question")]
)
context_based_evals.as_dataframe()
```

What happened in this code:

We create an Evidently dataset object.

Simultaneously, we add descriptors: evaluators that score each row.

We use a built-in LLM judge metric `ContextQualityLLMEval`.

i You can also choose a different evaluator LLM or modify the prompt. See [LLM judge parameters](#).

Here is what you get:

	Question	Context	Response	ContextQuality	ContextQuality reasoning
0	Why do flowers bloom in spring?	Plants require extra care during cold months. You should keep them indoors.	because of the rising temperatures	INVALID	The text does not provide any information about why flowers bloom in spring. Instead, it discusses care for plants during cold months, which is unrelated to the blooming of flowers.
1	Why do we yawn when we see someone else yawn?	Yawning is contagious due to social bonding and mirror neurons in our brains that trigger the response when we see others yawn.	because it's a glitch in the matrix	VALID	The text provides sufficient information to answer the question by explaining that yawning is contagious due to social bonding and mirror neurons, which triggers the response when observing others yawn.
2	How far is Saturn from Earth?	The distance between Earth and Saturn varies, but on average, Saturn is about 1.4 billion kilometers (886 million miles) away from Earth.	about 1.4 billion kilometers	VALID	The text provides a clear answer to the question by stating that the average distance from Earth to Saturn is about 1.4 billion kilometers (886 million miles). This information directly addresses the question asked.
3	Where do penguins live?	Penguins primarily live in the Southern Hemisphere, with most species found in Antarctica, as well as on islands and coastlines of South America, Africa, Australia, and New Zealand.	mostly in Antarctica and southern regions	VALID	The text provides clear and sufficient information about where penguins live, specifically mentioning the Southern Hemisphere, Antarctica, and various regions in South America, Africa, Australia, and New Zealand.



≡ LLM Tutorials > RAG evals

```
context_based_evals = Dataset.from_pandas(
    synthetic_df,
    data_definition=DataDefinition(text_columns=["Question", "Context", "Response"],
    descriptors=[ContextRelevance("Question", "Context",
                                output_scores=True,
                                aggregation_method="hit",
                                method="llm",
                                alias="Hit")])
)
context_based_evals.as_dataframe()
```

In this case you will get a binary “Hit” on whether the context is relevant or not.

	Question	Context	Response	Hit	Hit scores
0	Why do flowers bloom in spring?	Plants require extra care during cold months. You should keep them indoors.	because of the rising temperatures	0	[0.0]
1	Why do we yawn when we see someone else yawn?	Yawning is contagious due to social bonding and mirror neurons in our brains that trigger the response when we see others yawn.	because it's a glitch in the matrix	1	[1.0]
2	How far is Saturn from Earth?	The distance between Earth and Saturn varies, but on average, Saturn is about 1.4 billion kilometers (886 million miles) away from Earth.	about 1.4 billion kilometers	1	[1.0]
3	Where do penguins live?	Penguins primarily live in the Southern Hemisphere, with most species found in Antarctica, as well as on islands and coastlines of South America, Africa, Australia, and New Zealand.	mostly in Antarctica and southern regions	1	[1.0]

It's more useful for multiple context, though.

Multiple Contexts

RAG systems often retrieve multiple chunks. In this case, we can assess the relevance of each individual chunk first.

Let's generate a toy dataset. Pass all contexts as a list.

```
synthetic_data = [
    ["Why are bananas healthy?", ["Bananas are rich in potassium.", "Bananas provide many health benefits."], "Bananas are healthy because they are rich in potassium and provide many health benefits."],
    ["How do you cook potatoes?", ["Potatoes are easy to grow.", "The best way to cook potatoes is to boil them."], "Potatoes are easy to grow and the best way to cook them is to boil them."],
]
columns = ["Question", "Context", "Response"]
synthetic_df_2 = pd.DataFrame(synthetic_data, columns=columns)
```

Hit Rate. To aggregate the results per query, we can assess if at least one retrieved chunk



≡ LLM Tutorials > RAG evals

```
data_definition=DataDefinition(text_columns=["Question", "Context", "Response"]
descriptors=[ContextRelevance("Question", "Context",
                                output_scores=True,
                                aggregation_method="hit",
                                method="llm",
                                alias="Hit")]
)
context_based_evals.as_dataframe()
```

You can see the list of individual relevance scores that appear in the same order as your chunks.

	Question	Context	Response	Hit	Hit scores
0	Why are bananas healthy?	[Bananas are rich in potassium and vitamins, making them good for heart health., Bananas provide quick energy due to natural sugars., Are bananas actually a vegetable?]	because they are rich in nutrients	1	[1.0, 1.0, 0.0]
1	How do you cook potatoes?	[Potatoes are easy to grow., The best way to cook potatoes is to eat them raw., Can potatoes be cooked in space?]	boil, bake, or fry them	0	[0.0, 0.1, 0.1]

Mean Relevance. Alternatively, you can compute an average relevance score.

```
context_based_evals = Dataset.from_pandas(
    synthetic_df_2,
    data_definition=DataDefinition(text_columns=["Question", "Context", "Response"]
    descriptors=[ContextRelevance("Question", "Context",
                                    output_scores=True,
                                    aggregation_method="mean",
                                    method="llm",
                                    alias="Relevance")]
    )
context_based_evals.as_dataframe()
```

Here is an example result:

	Question	Context	Response	Relevance	Relevance scores
0	Why are bananas healthy?	[Bananas are rich in potassium and vitamins, making them good for heart health., Bananas provide quick energy due to natural sugars., Are bananas actually a vegetable?]	because they are rich in nutrients	0.666667	[1.0, 1.0, 0.0]
1	How do you cook potatoes?	[Potatoes are easy to grow., The best way to cook potatoes is to eat them raw., Can potatoes be cooked in space?]	boil, bake, or fry them	0.033333	[0.0, 0.0, 0.1]

known correct answers.

❗ **Synthetic data.** You can generate a ground truth dataset for your RAG using [Evidently Platform](#).

Let's generate a new toy example with "target" column:

```
synthetic_data = [
    ["Why do we yawn?", "because it's a glitch in the matrix", "Due to mirror neur
    ["Why do flowers bloom?", "Because of rising temperatures", "Because it gets w
]
columns = ["Question", "Response", "Target"]
synthetic_df_3 = pd.DataFrame(synthetic_data, columns=columns)
```

There are multiple ways to run this comparison, including LLM-based matching (`CorrectnessLLMEval`) and non-LLM methods like Semantic similarity and BERTScore. Let's run all three at once, but we'd recommend choosing the one:

```
context_based_evals = Dataset.from_pandas(
    synthetic_df_3,
    data_definition=DataDefinition(text_columns=["Question", "Response", "Target"]
    descriptors=[
        CorrectnessLLMEval("Response", target_output="Target"),
        BERTScore(columns=["Response", "Target"], alias="BERTScore"),
        SemanticSimilarity(columns=["Response", "Target"], alias="Semantic Similar
    ]
)
context_based_evals.as_dataframe()
```

Here is what you get:

	Question	Response	Target	Correctness	Correctness reasoning	BERTScore	Semantic Similarity
0	Why do we yawn when we see someone else yawn?	because it's a glitch in the matrix.	Due to social bonding and mirror neurons in our brains.	INCORRECT	The OUTPUT introduces a phrase 'it's a glitch in the matrix' which does not convey the original meaning of social bonding and mirror neurons presented in the REFERENCE. It alters the key details and does not align with the factual basis of the REFERENCE.	0.544553	0.640252
1	Why do flowers bloom in spring?	Because of the the rising temperatures.	Because it is getting warmer.	CORRECT	The OUTPUT conveys the same underlying meaning as the REFERENCE by indicating that it is getting warmer due to rising temperatures. The different wording does not change the original idea.	0.730837	0.889425
2	Why are bananas healthy?	Because they are rich in nutrients.	Because they contain a lot of nutrients.	CORRECT	The OUTPUT states that 'they are rich in nutrients,' which conveys the same meaning as the REFERENCE 'they contain a lot of nutrients.' Both sentences imply a high nutrient content, preserving the original meaning.	0.878887	0.983975

Without Ground Truth

If you don't have reference answers, you can use reference-free LLM judges to assess response quality. For example, here is you how can run evaluation for **Faithfulness** to detect if the response is contradictory or unfaithful to the context:

```
context_based_evals = Dataset.from_pandas(  
    synthetic_df,  
    data_definition=DataDefinition(text_columns=["Question", "Context", "Response"],  
    descriptors=[FaithfulnessLLMEval("Response", context="Context")]  
)  
context_based_evals.as_dataframe()
```

Here is an example result:

	Question	Context	Response	Faithfulness	Faithfulness reasoning
0	Why do flowers bloom in spring?	Plants require extra care during cold months. You should keep them indoors.	because of the rising temperatures	UNFAITHFUL	The text 'because of the rising temperatures' contradicts the source which states that plants require extra care during cold months and should be kept indoors. Rising temperatures suggest warmer weather, which is contrary to the need for extra care during cold months.
1	Why do we yawn when we see someone else yawn?	Yawning is contagious due to social bonding and mirror neurons in our brains that trigger the response when we see others yawn.	because it's a glitch in the matrix	UNFAITHFUL	The response 'because it's a glitch in the matrix' does not relate to the information provided in the source about yawning being contagious due to social bonding and mirror neurons. It contradicts the source by introducing an unrelated concept without any grounding in the original source material.
2	How far is Saturn from Earth?	The distance between Earth and Saturn varies, but on average, Saturn is about 1.4 billion kilometers (886 million miles) away from Earth.	about 1.4 billion kilometers	FAITHFUL	The text accurately references the distance to Saturn as being about 1.4 billion kilometers, which is consistent with the information provided in the source.
3	Where do penguins live?	Penguins primarily live in the Southern Hemisphere, with most species found in Antarctica, as well as on islands and coastlines of South America, Africa, Australia, and New Zealand.	mostly in Antarctica and southern regions	FAITHFUL	The text accurately reflects information from the source by indicating that penguins primarily live in Antarctica and southern regions, which aligns with the description that they are mostly found in the Southern Hemisphere and particularly in Antarctica.

You can add other useful checks over your final response like:

Length constraints: are responses within expected limits?

Refusal rate: monitoring how often the system declines questions.

String matching: checking for required wording (e.g., disclaimers).

Response tone: ensuring responses match the intended style.

❗ Available evaluators. [Check a full list of available descriptors.](#)

4. Get Reports



≡ LLM Tutorials > RAG evals

Score data. Once you have a pandas dataframe `synthetic_df`, you create an Evidently dataset object and choose the selected descriptors by simply listing them.

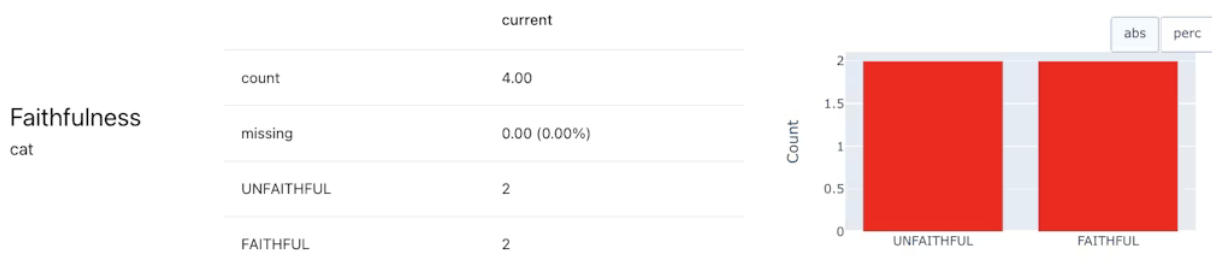
```
context_based_evals = Dataset.from_pandas(
    synthetic_df,
    data_definition=DataDefinition(
        text_columns=["Question", "Context", "Response"],
    ),
    descriptors=[
        FaithfulnessLLEval("Response", context="Context"),
        ContextQualityLLEval("Context", question="Question"),
    ]
)
# context_based_evals.as_dataframe()
```

Get a Report. Instead of rendering the results as a dataframe, you create a [Report](#).

```
report = Report([
    TextEvals()
])

my_eval = report.run(context_based_evals, None)
my_eval
```

This will render an HTML report in the notebook cell. You can use other [export options](#), like `as_dict()` for a Python dictionary output.





≡ LLM Tutorials > RAG evals



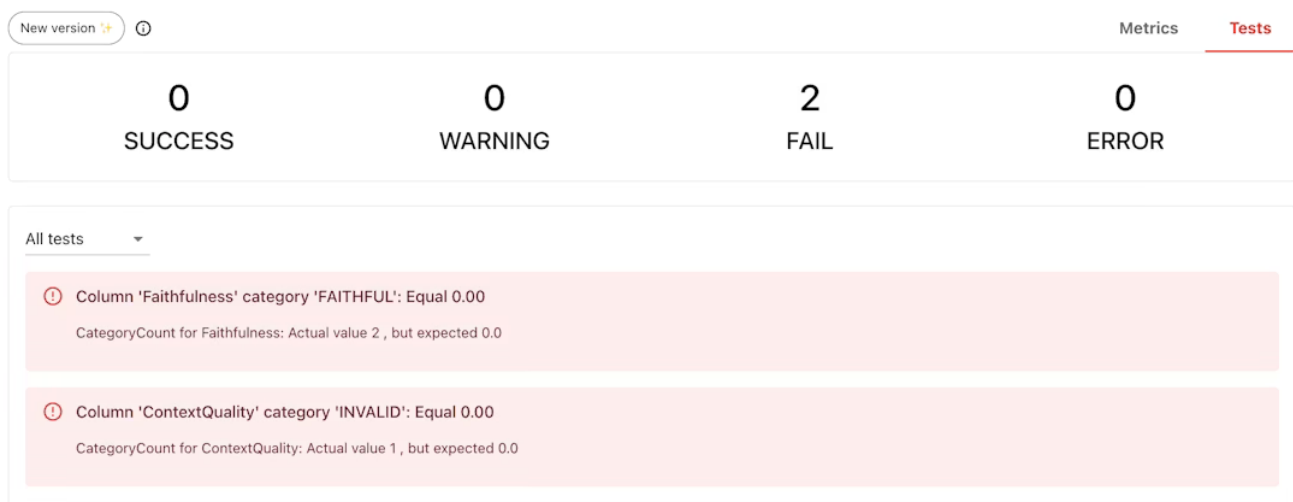
This lets you see a well-rounded evaluation. In this toy example, we can see that the system generally retrieves the right data well but struggles with generation. The next step could be improving your prompt to ensure responses stay true to context.

Add test conditions. You can also set up explicit pass/fail tests based on expected score distributions using the Tests. These are conditional expectations you add to metrics.

```
report = Report([
    TextEvals(),
    CategoryCount(column="Faithfulness", category="UNFAITHFUL", tests=[eq(0)]),
    CategoryCount(column="ContextQuality", category="INVALID", tests=[eq(0)])
])

my_eval = report.run(context_based_evals, None)
my_eval
```

In this case, we expect all retrieved contexts to be valid and all responses to be faithful, so our tests fail. You can adjust these conditions — for example, allowing a certain percentage of responses to fail.





≡ LLM Tutorials > RAG evals

Set up Evidently Cloud

Sign up for a free [Evidently Cloud account](#).

Create an **Organization** if you log in for the first time. Get an ID of your organization. ([Link](#)).

Get an **API token**. Click the **Key** icon in the left menu. Generate and save the token. ([Link](#)).

Import the components to connect with Evidently Cloud:

```
from evidently.ui.workspace import CloudWorkspace
```



Create a Project

Connect to Evidently Cloud using your API token:

```
ws = CloudWorkspace(token="YOUR_API_TOKEN", url="https://app.evidently
```



Create a Project within your Organization, or connect to an existing Project:

```
project = ws.create_project("My project name", org_id="YOUR_ORG_ID")
project.description = "My project description"
project.save()

# or project = ws.get_project("PROJECT_ID")
```



Alternatively, retrieve an existing project:

```
# project = ws.get_project("PROJECT_ID")
```



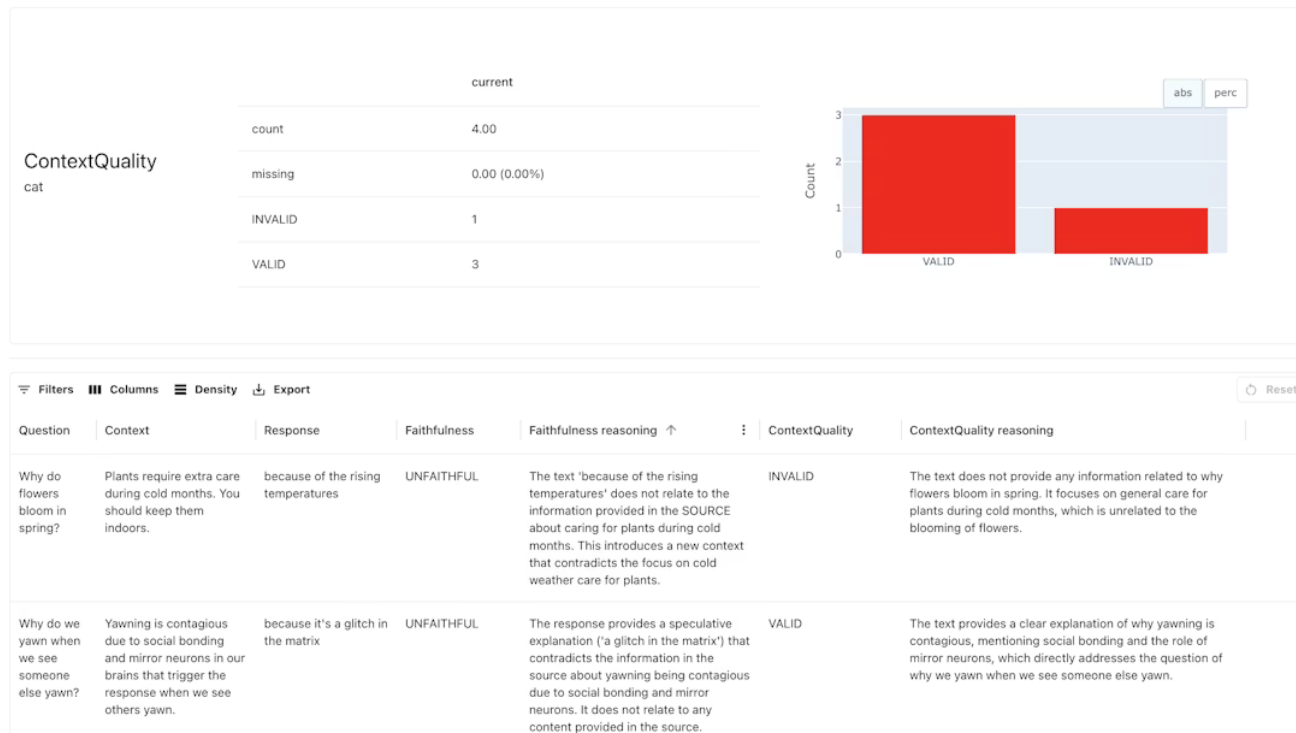
Send your eval

Since you already created the eval, you can simply upload it to the Evidently Cloud.



LLM Tutorials > RAG evals

data that's easy to interact with.



FiltersColumnsDensityExportReset

Question	Context	Response	Faithfulness	Faithfulness reasoning ↑	ContextQuality	ContextQuality reasoning
Why do flowers bloom in spring?	Plants require extra care during cold months. You should keep them indoors.	because of the rising temperatures	UNFAITHFUL	The text 'because of the rising temperatures' does not relate to the information provided in the SOURCE about caring for plants during cold months. This introduces a new context that contradicts the focus on cold weather care for plants.	INVALID	The text does not provide any information related to why flowers bloom in spring. It focuses on general care for plants during cold months, which is unrelated to the blooming of flowers.
Why do we yawn when we see someone else yawn?	Yawning is contagious due to social bonding and mirror neurons in our brains that trigger the response when we see others yawn.	because it's a glitch in the matrix	UNFAITHFUL	The response provides a speculative explanation ('a glitch in the matrix') that contradicts the information in the source about yawning being contagious due to social bonding and mirror neurons. It does not relate to any content provided in the source.	VALID	The text provides a clear explanation of why yawning is contagious, mentioning social bonding and the role of mirror neurons, which directly addresses the question of why we yawn when we see someone else yawn.

What's Next?

Considering implementing a regression testing at every update to monitor how your RAG system retrieval and response quality changes.

< LLM evaluations

LLM as a judge >

Powered by **mintlify**



☰ LLM Tutorials > RAG evals
