

# Beginner's Guide to RAG Evaluation

AI by Hand 

Prof. Tom Yeh



University of Colorado  
Boulder

Hosted by:



# Poll

Beginner's Guide to RAG Evaluation - AI by Hand 📝



University of Colorado  
Boulder

Download: <https://by-hand.ai/rageval>

# What do you want to learn about RAG Evaluation?

You can see how people vote. [Learn more](#)

Tools ✓

39%

Datasets ✓

8%

Metrics ✓

50%

Other (Please Comment) ✓

3%

265 votes • Poll closed

# RAG

Beginner's Guide to RAG Evaluation - AI by Hand 📝



University of Colorado  
Boulder

Download: <https://by-hand.ai/rageval>

# RAG

Retriever

## User Question

What pets?

Answer

Augmentation

Generator

# Trace

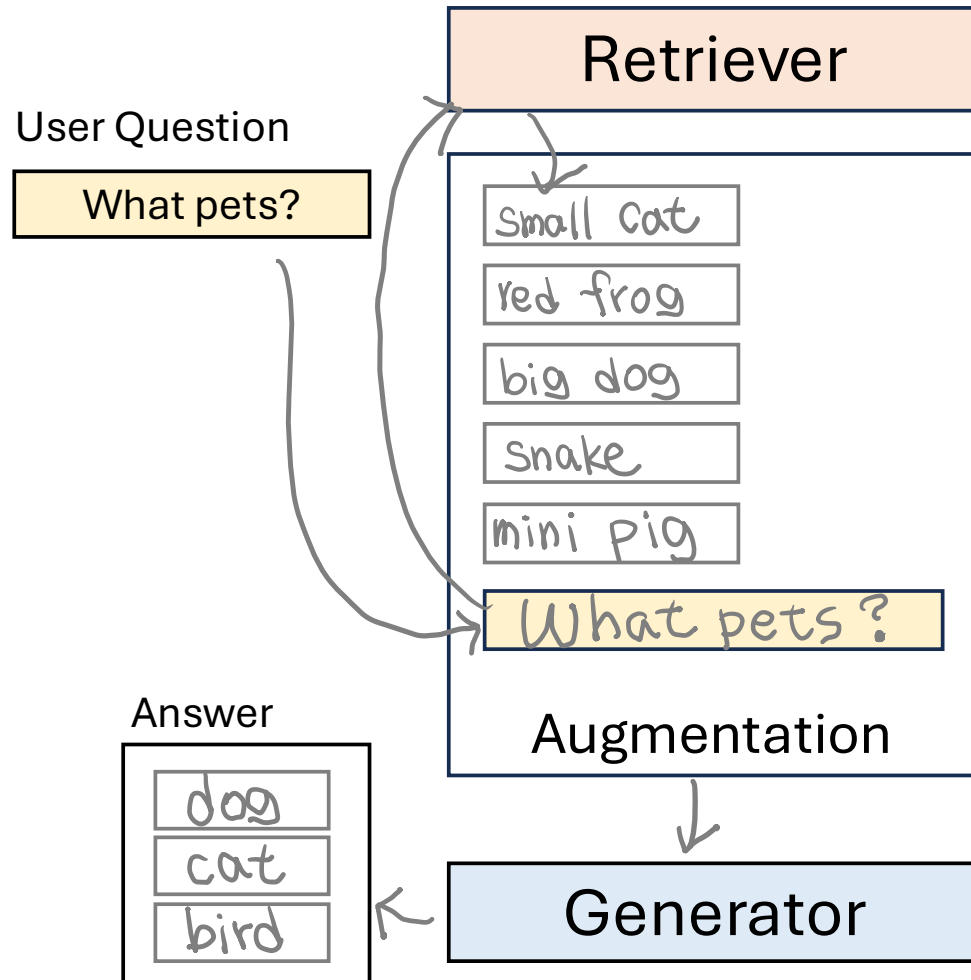
Beginner's Guide to RAG Evaluation - AI by Hand 📝



University of Colorado  
Boulder

Download: <https://by-hand.ai/rageval>

# RAG



# Trace

Retrieval


Prompt Embedding


Vector Store


Context Encode

Generation

 **langfuse** Public

 Watch 21 ▾

 Fork 495 ▾

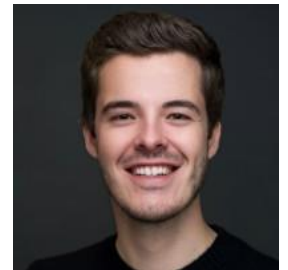
 Star 5.4k ▾

# Langfuse

Beginner's Guide to RAG Evaluation - AI by Hand 📌



Special Thanks

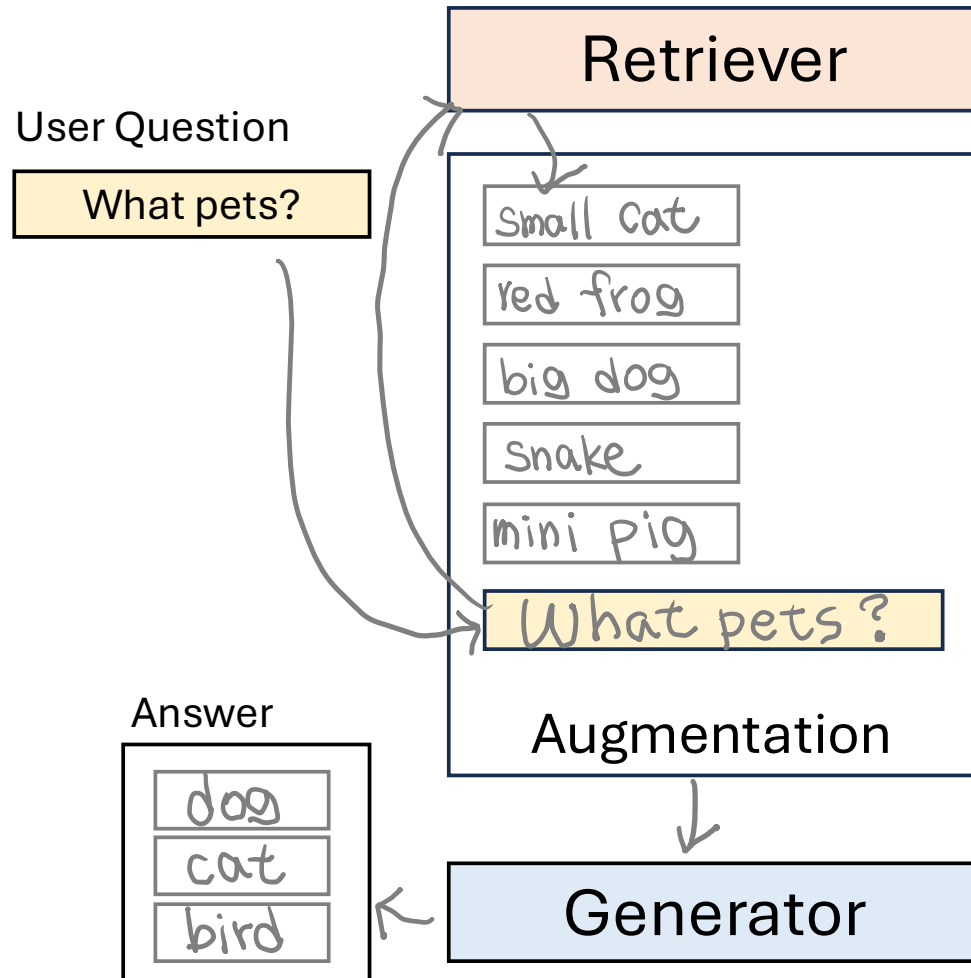


Marc Kling  
Co-founder

Download: <https://by-hand.ai/rageval>



# RAG



## Trace Timeline ⊕ ⊖

- TRACE qa
- SPAN retrieval
  - GENERATION prompt-embedding
  - SPAN vector-store
  - SPAN context-encoding
  - SPAN fetch-prompt-from-langfuse
  - GENERATION generation



👋 Do you have any questions about Langfuse? Ask me!

⚠️ *Warning: Do not enter sensitive information. All chat messages can be viewed in the public demo project. Responses may be inaccurate. Please check the documentation for details or reach out to us via the chat widget.*

Send a message





👋 Do you have any questions about Langfuse? Ask me!

⚠️ *Warning: Do not enter sensitive information. All chat messages can be viewed in the public demo project. Responses may be inaccurate. Please check the documentation for details or reach out to us via the chat widget.*



What are good pets?



Sorry, I don't know how to help with that.

Send a message





👋 Do you have any questions about Langfuse? Ask me!

⚠️ *Warning: Do not enter sensitive information. All chat messages can be viewed in the public demo project. Responses may be inaccurate. Please check the documentation for details or reach out to us via the chat widget.*



What are good pets?



Sorry, I don't know how to help with that.

Send a message





👋 Do you have any questions about Langfuse? Ask me!

⚠️ *Warning: Do not enter sensitive information. All chat messages can be viewed in the public demo project. Responses may be inaccurate. Please check the documentation for details or reach out to us via the chat widget.*

---



What are good pets?

---



Sorry, I don't know how to help with that.

---



What is a trace?

---



A trace in Langfuse represents a single request or operation within your application. It includes the following components:

- **Overall Input and Output:** This represents the complete data processed during that request.
- **Metadata:** Details about the request such as the user, the session, and any tags associated with it.

Send a message



Search by id, name, user id

Q

Filter

24 hours

(7/29)

<input type="checkbox"/>	ID	User <sup>Ⓢ</sup>	Usage	# conciseness-v1 (eval)	# helpfulness-1 (eval)	# toxicity-v2 (eval)	Total Cost	Tags <sup>Ⓢ</sup>	Action
<input type="checkbox"/>	☆ 7c973011-43...	u-SoAo6Be	943 → 212 (Σ 1,155)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0003	with-context	
<input type="checkbox"/>	☆ fc8a6acd-7d...	u-SoAo6Be	99 → 11 (Σ 110)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ 3e77f6c2-74...	u-9qRwS0z	159 → 11 (Σ 170)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ eebb663a-a0...	u-9qRwS0z	136 → 11 (Σ 147)	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ be3e1a0f-1b...	u-9qRwS0z	108 → 11 (Σ 119)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ 9e09f16c-51...	u-K8t215e	148 → 11 (Σ 159)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ c7fcfb43-d9...	u-K8t215e	740 → 11 (Σ 751)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0001	with-context	
<input type="checkbox"/>	☆ 16f76d22-17...	u-K8t215e	358 → 11 (Σ 369)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0001	with-context	
<input type="checkbox"/>	☆ f31a7481-e4f...	u-2PkSSae	97 → 30 (Σ 127)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ 0b76d3e9-b...	u-0ojLo4C	483 → 36 (Σ 519)	0.1000 <sup>Ⓢ</sup>	0.2000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0001	with-context	
<input type="checkbox"/>	☆ 75aecee5-04...	u-0ojLo4C	389 → 32 (Σ 421)	0.0000 <sup>Ⓢ</sup>	0.2000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0001	with-context	
<input type="checkbox"/>	☆ d7c311a6-eb...	u-Ux4V8EW	4,014 → 729 (Σ 4,743)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.001	with-context	
<input type="checkbox"/>	☆ 19acd92c-78...	u-Ux4V8EW	3,001 → 398 (Σ 3,399)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0007	with-context	
<input type="checkbox"/>	☆ 81d0b569-6...	u-Ux4V8EW	2,244 → 333 (Σ 2,577)	0.9000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0005	with-context	
<input type="checkbox"/>	☆ 07e35225-67...	u-Ux4V8EW	2,845 → 390 (Σ 3,235)	0.4000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0007	with-context	
<input type="checkbox"/>	☆ cca2662c-bc...	u-Ux4V8EW	1,548 → 464 (Σ 2,012)	0.1000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0005	with-context	
<input type="checkbox"/>	☆ 37ff7b0b-6a...	u-Ux4V8EW	1,577 → 56 (Σ 1,633)	0.1000 <sup>Ⓢ</sup>	0.9000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0003	with-context	
<input type="checkbox"/>	☆ d0948efe-d1...	u-Ux4V8EW	864 → 11 (Σ 875)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0001	no-context	
<input type="checkbox"/>	☆ b1467a5a-fd...	u-Ux4V8EW	1,831 → 348 (Σ 2,179)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0005	with-context	
<input type="checkbox"/>	☆ 269591b4-e...	u-Ux4V8EW	1,510 → 365 (Σ 1,875)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0004	with-context	
<input type="checkbox"/>	☆ bbce4c6c-84...	u-GtpTQrm	870 → 232 (Σ 1,102)	0.9000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0003	with-context	
<input type="checkbox"/>	☆ 5bc36c39-d...	u-B6kKyPu	886 → 276 (Σ 1,162)	0.2000 <sup>Ⓢ</sup>	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0003	with-context	
<input type="checkbox"/>	☆ 0fb2c51c-d9...	u-oC3aYkF	504 → 40 (Σ 544)	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0001	with-context	
<input type="checkbox"/>	☆ d72b106b-11...	u-Hmj2f7d	89 → 11 (Σ 100)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ 77317131-50f...	u-0b6oEKx	91 → 11 (Σ 102)	1.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.00	no-context	
<input type="checkbox"/>	☆ 26efa5c0-04...	u-pX1dxFk	1,298 → 392 (Σ 1,690)	0.9000 <sup>Ⓢ</sup>	0.9000 <sup>Ⓢ</sup>	0.0000 <sup>Ⓢ</sup>	\$0.0004	with-context	

Preview

Scores

TRACE

qa

8/28/2024, 8:22:27 AM

1.69s

99 → 11 (Σ 110)

Release: 362b43416c0c1fed137daf89eed8e53141282b17

Annotate

Add to dataset

Input

"What are good pets?"

Output

"Sorry, I don't know how to help with that."

Metadata

```
{  
  pathname: "/docs/demo"  
}
```

Scores

EVAL

conciseness-v1: 1.00

contextrelevance-v1: 0.00

correctness-v1: 0.10

hallucination-v2: 0.00

helpfulness-1: 0.00

language-detector-v3: 0.00

thank-you-v1: 0.00

toxicity-v2: 0.00

TRACE

qa

1.69s

conciseness-v1: 1.00

contextrelevance-v1: 0.00

correctness-v1: 0.10

hallucination-v2: 0.00

helpfulness-1: 0.00

language-detector-v3: 0.00

thank-you-v1: 0.00

toxicity-v2: 0.00

SPAN

retrieval

0.97s

WARNING

GENERATION

prompt-embedding

0.21s 5 → 0 (Σ 5)

DEBUG

SPAN

vector-store

0.76s

SPAN

context-encoding

0.00s

SPAN

fetch-prompt-from-langfuse

0.24s

GENERATION

generation

0.47s 94 → 11 (Σ 105)

WARNING

“What are good pets?”

Preview

Scores

GENERATION

prompt-embedding

8/28/2024, 8:22:27 AM

Latency: 0.21s

5 prompt → 0 completion (Σ 5)

text-embedding-ada-002

\$0.00

Annotate

>\_ Test in playground

Add to dataset

Input

"What are good pets?"

Output

null

## Trace Timeline

- TRACE qa
  - SPAN retrieval
    - GENERATION prompt-embedding
    - SPAN vector-store
    - SPAN context-encoding
    - SPAN fetch-prompt-from-langfuse
    - GENERATION generation

“What is a trace?”

Preview

Scores

GENERATION

prompt-embedding

8/28/2024, 8:22:32 AM

Latency: 0.22s

5 prompt → 0 completion (Σ 5)

text-embedding-ada-002

\$0.00

Annotate

>\_ Test in playground

Add to dataset

Input

"What is a trace?"

Output

null



“What are good pets?”

PreviewScores

SPANvector-store

8/28/2024, 8:22:27 AM

Latency: 0.76s

AnnotateAdd to dataset

Input

[0 ... 99]

[100 ... 199]

[200 ... 299]

[300 ... 399]

[400 ... 499]

[500 ... 599]

[600 ... 699]

[700 ... 799]

[800 ... 899]

[900 ... 999]

[1000 ... 1099]

[1100 ... 1199]

[1200 ... 1299]

[1300 ... 1399]

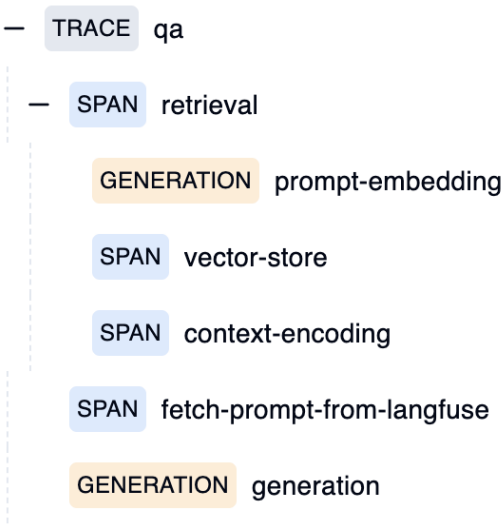
[1400 ... 1499]

[1500 ... 1535]

Output

[0 ... 99]

Trace Timeline



“What is a trace?”

PreviewScores

SPANvector-store

8/28/2024, 8:22:32 AM

Latency: 0.51s

AnnotateAdd to dataset

Input

[0 ... 99]

[100 ... 199]

[200 ... 299]

[300 ... 399]

[400 ... 499]

[500 ... 599]

[600 ... 699]

[700 ... 799]

[800 ... 899]

[900 ... 999]

[1000 ... 1099]

[1100 ... 1199]

[1200 ... 1299]

[1300 ... 1399]

[1400 ... 1499]

[1500 ... 1535]

Output

0: {slug: "description-each-trace-has-a-unique-url-that-you-can-use-to-share-it-with-others-or-to-access-it-directly", content: "## description: Each trace has a unique URL that you can use to share it with others or to access it directly.", heading: "description: Each trace has a unique URL that you can use to share it with others or to access it directly.", page\_id: 20662, similarity: 0.851839892699189}

1: {slug: "trace-urls", content: "# Trace URLs", heading: "Trace URLs", page\_id: 20662, similarity: 0.842895128049308}

Each trace has a unique URL that you can use to share it with others or to access it directly.

“What are good pets?”

## Trace Timeline

“What is a trace?”

Preview

Scores

GENERATION

generation

8/28/2024, 8:22:28 AM

Prompt: qa-answer-no-context-chat - v2


Time to first token: 0.35s

Latency: 0.47s

94 prompt → 11 completion (Σ 105)


gpt-4o-mini


\$0.00





Annotate

>\_ Test in playground



Add to dataset 

Pretty 



JSON

system  


You are a very enthusiastic Langfuse representative who loves to help people! Langfuse is an open-source observability tool for developers of applications that use Large Language Models (LLMs).  
Answer with "Sorry, I don't know how to help with that." if the question is not related to Langfuse or if you are unable to answer it based on the context.  
be nice!

user  

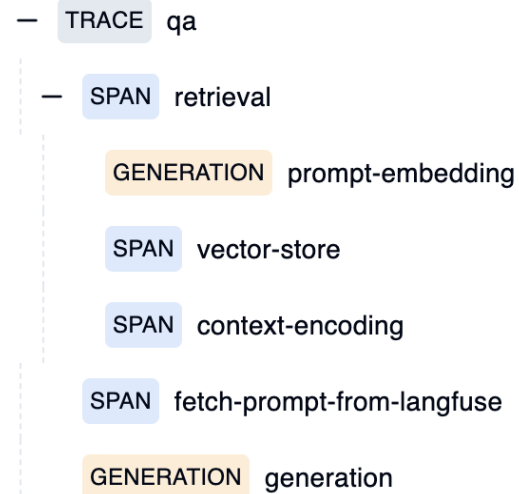
What are good pets?

assistant  

Sorry, I don't know how to help with that.

Status Message 

"Refused to answer"



Preview

Scores

GENERATION

generation

8/28/2024, 8:22:33 AM

Prompt: qa-answer-with-context-chat - v12


Time to first token: 0.50s

Latency: 8.29s

938 prompt → 212 completion (Σ 1150)


gpt-4o-mini


\$0.0003





Annotate

>\_ Test in playground

Add to dataset 



Pretty 

JSON



system  

You are a very enthusiastic Langfuse representative who loves to help people! Langfuse is an open-source observability tool for developers of applications that use Large Language Models (LLMs). Given the following sections from the Langfuse documentation, answer the question using only that information, outputted in markdown format. Refer to the respective links of the documentation and be nice. Be more kind and do not be rude. Include two emojis at the end.

Show 5 more ...

user  

What is a trace?

assistant  

A trace in Langfuse is a comprehensive representation of a single request or operation. It consists of the following key components:

- **Trace Object:** This typically contains overall input and output of the function, alongside metadata about the request, like the user and session information. You can also add tags to help categorize and filter traces easily.


"What are good pets?"


## Trace Timeline


"What is a trace?"


Preview


Scores


conciseness-v1: 1.00 


contextrelevance-v1: 0.00 


correctness-v1: 0.10 

hallucination-v2: 0.00 

helpfulness-1: 0.00 

language-detector-v3: 0.00 

thank-you-v1: 0.00 

toxicity-v2: 0.00 

user

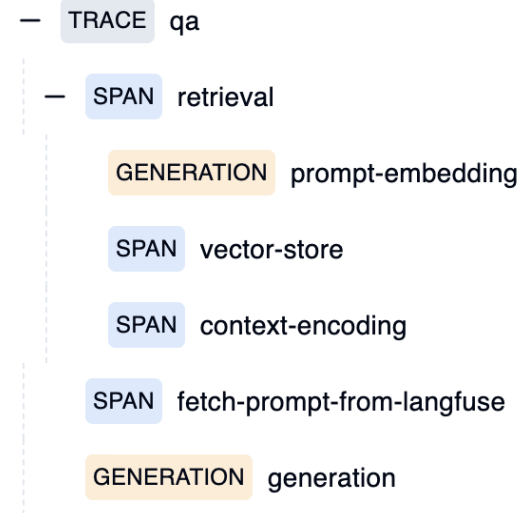
What are good pets?

assistant

Sorry, I don't know how to help with that.

Status Message

"Refused to answer"



Show 5 more ...

user

What is a trace?

assistant

A trace in Langfuse is a comprehensive representation of a single request or operation. It consists of the following key components:

- **Trace Object:** This typically contains overall input and output of the function, alongside metadata about the request, like the user and session information. You can also add tags to help categorize and filter traces easily.

# Conciseness

Evaluate the                      of the generation on a continuous scale from 0 to 1. A generation can be considered concise (Score: 1) if it directly and succinctly answers the question posed, focusing specifically on the information requested without including unnecessary, irrelevant, or excessive details.

# conciseness

## Prompt

Evaluate the conciseness of the generation on a continuous scale from 0 to 1. A generation can be considered concise (Score: 1) if it directly and succinctly answers the question posed, focusing specifically on the information requested without including unnecessary, irrelevant, or excessive details.

Example:  
Query: Can eating carrots improve your vision?  
Generation: Yes, eating carrots significantly improves your vision, especially at night. This is why people who eat lots of carrots never need glasses. Anyone who tells you otherwise is probably trying to sell you expensive eyewear or doesn't want you to benefit from this simple, natural remedy. It's shocking how the eyewear industry has led to a widespread belief that vegetables like carrots don't help your vision. People are so gullible to fall for these money-making schemes.  
Score: 0.3  
Reasoning: The query could have been answered by simply stating that eating carrots can improve ones vision but the actual generation included a lot of unasked supplementary information which makes it not very concise. However, if present, a scientific

You can use `{{variable}}` to insert variables into your prompt. **Note:** Variables must be alphabetical characters or underscores. The following variables are available:

query

generation

## Score

provide a score between 0 and 1

We use function calls to extract data from the LLM. Specify what the LLM should return for the score.

## Reasoning

provide a one sentence reasoning

We use function calls to extract data from the LLM. Specify what the LLM should return for the reasoning.



v2 - 7/15/2024



## Model

### Provider

openai



### Model name

gpt-4o



### Temperature

0



### Output token limit

256



### Top P

1



### API key

...vTlu

The LLM API key is used for each execution and will incur costs.

# Integration



```
from llama_index.core import Settings

from llama_index.core.callbacks import CallbackManager

from langfuse.llama_index import LlamaIndexCallbackHandler

Settings.callback_manager = CallbackManager([LlamaIndexCallbackHandler()])
```



ragas

Public

Watch

34



Fork

606



Star

6.3k

# RAGAS – RAG Assessment

Beginner's Guide to RAG Evaluation - AI by Hand 📌



University of Colorado  
Boulder

Special Thanks



Shahul ES  
Founder

Download: <https://by-hand.ai/rageval>

# ragas score

## generation

### **faithfulness**

how factually accurate is  
the generated answer

### **answer relevancy**

how relevant is the generated  
answer to the question

## retrieval

### **context precision**

the signal to noise ratio of retrieved  
context

### **context recall**

can it retrieve all the relevant information  
required to answer the question



# ragas score

## generation

### **faithfulness**

how factually accurate is  
the generated answer

### **answer relevancy**

how relevant is the generated  
answer to the question

## retrieval

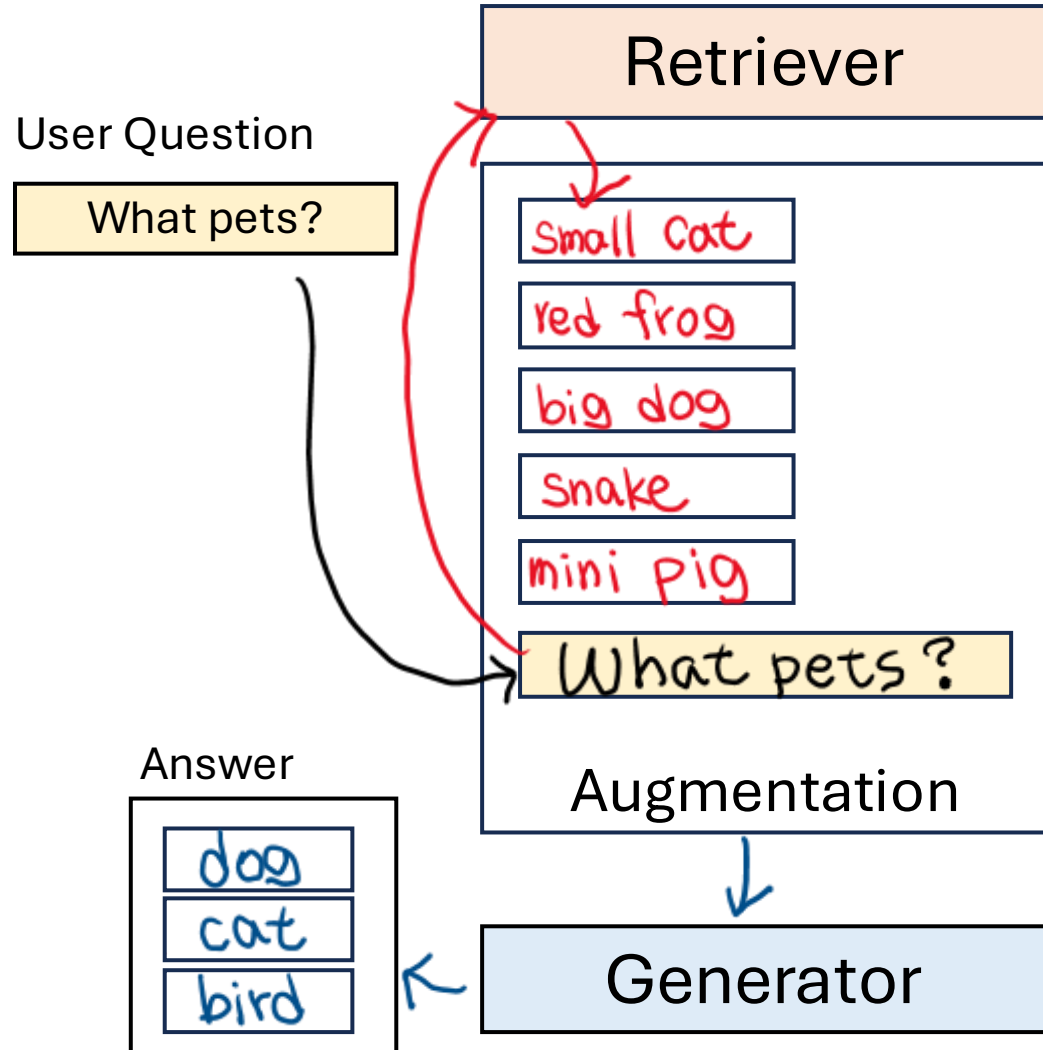
### **context precision**

the signal to noise ratio of retrieved  
context

### **context recall**

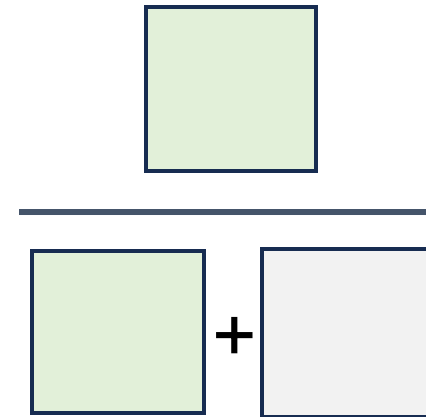
can it retrieve all the relevant information  
required to answer the question

# RAG

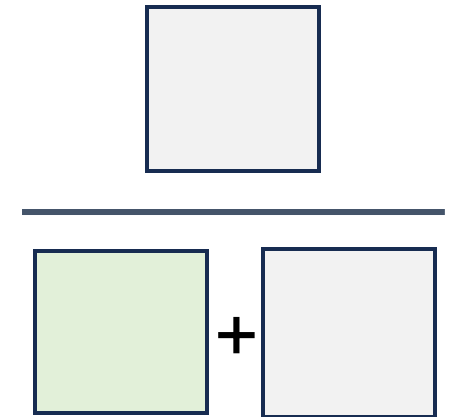


# Generator Metrics

Faithfulness



Hallucination



# Faithfulness

Your task is to judge the                      of a series of statements based on a given                     . For each statement you must return verdict as 1 if the statement can be directly inferred based on the context or 0 if the statement can not be directly inferred based on the context.

# ragas score

## generation

### faithfulness

how factually accurate is  
the generated answer

### answer relevancy

how relevant is the generated  
answer to the question

## retrieval

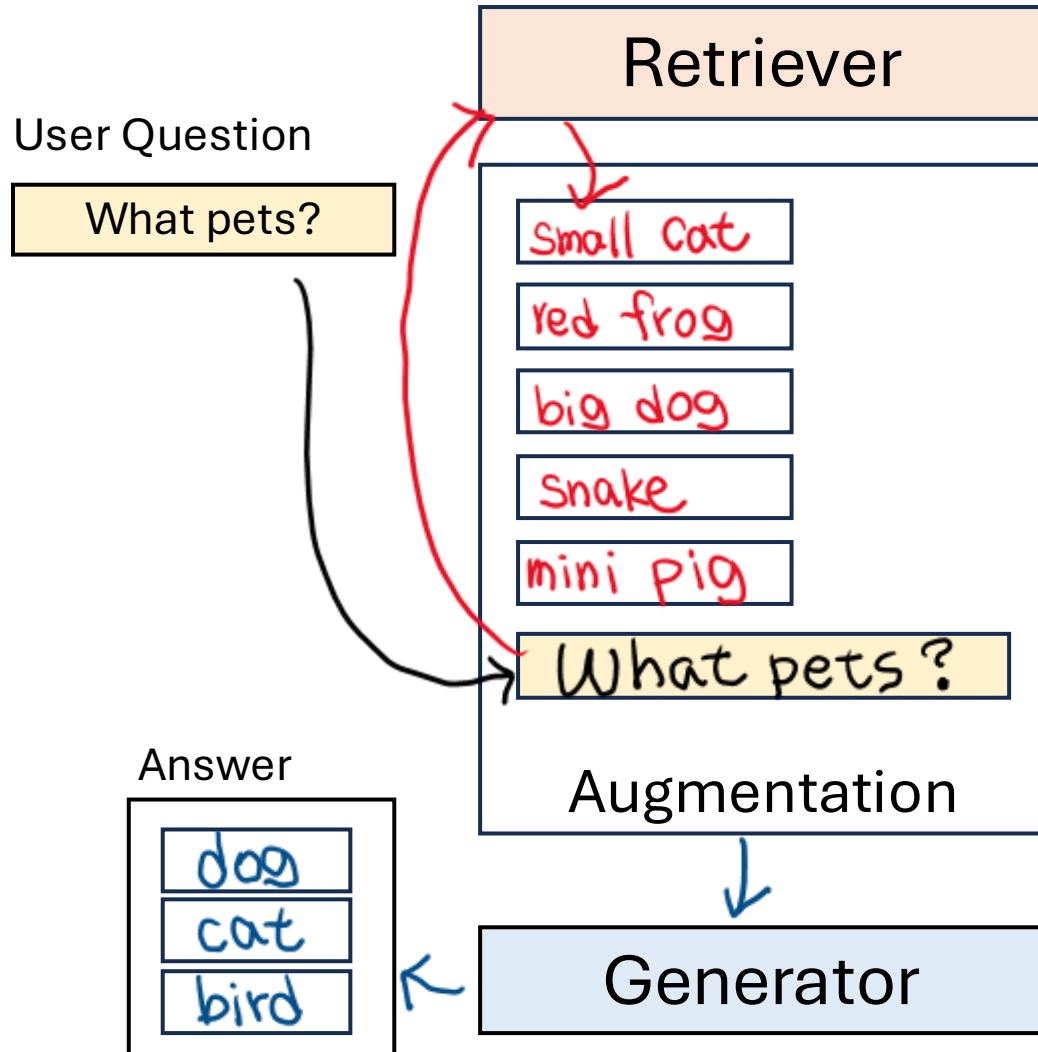
### context precision

the signal to noise ratio of retrieved  
context

### context recall

can it retrieve all the relevant information  
required to answer the question

# RAG



# Generator Metrics

## Answer Relevancy

	What pets?
What are common pets?	
What are domesticated animals?	
What are small animals?	

$$\sum \frac{1}{N}$$

\_\_\_\_\_

\_\_\_\_\_

# Question Generation

Generate a                      for the given answer and  
Identify if answer is                     . Give  
noncommittal as 1 if the answer is noncommittal  
and 0 if the answer is committal. A noncommittal  
answer is one that is evasive, vague, or  
ambiguous. For example, "I don't know" or "I'm  
not sure" are noncommittal answers"

# ragas score

## generation

### faithfulness

how factually accurate is  
the generated answer

### answer relevancy

how relevant is the generated  
answer to the question

## retrieval

### context precision

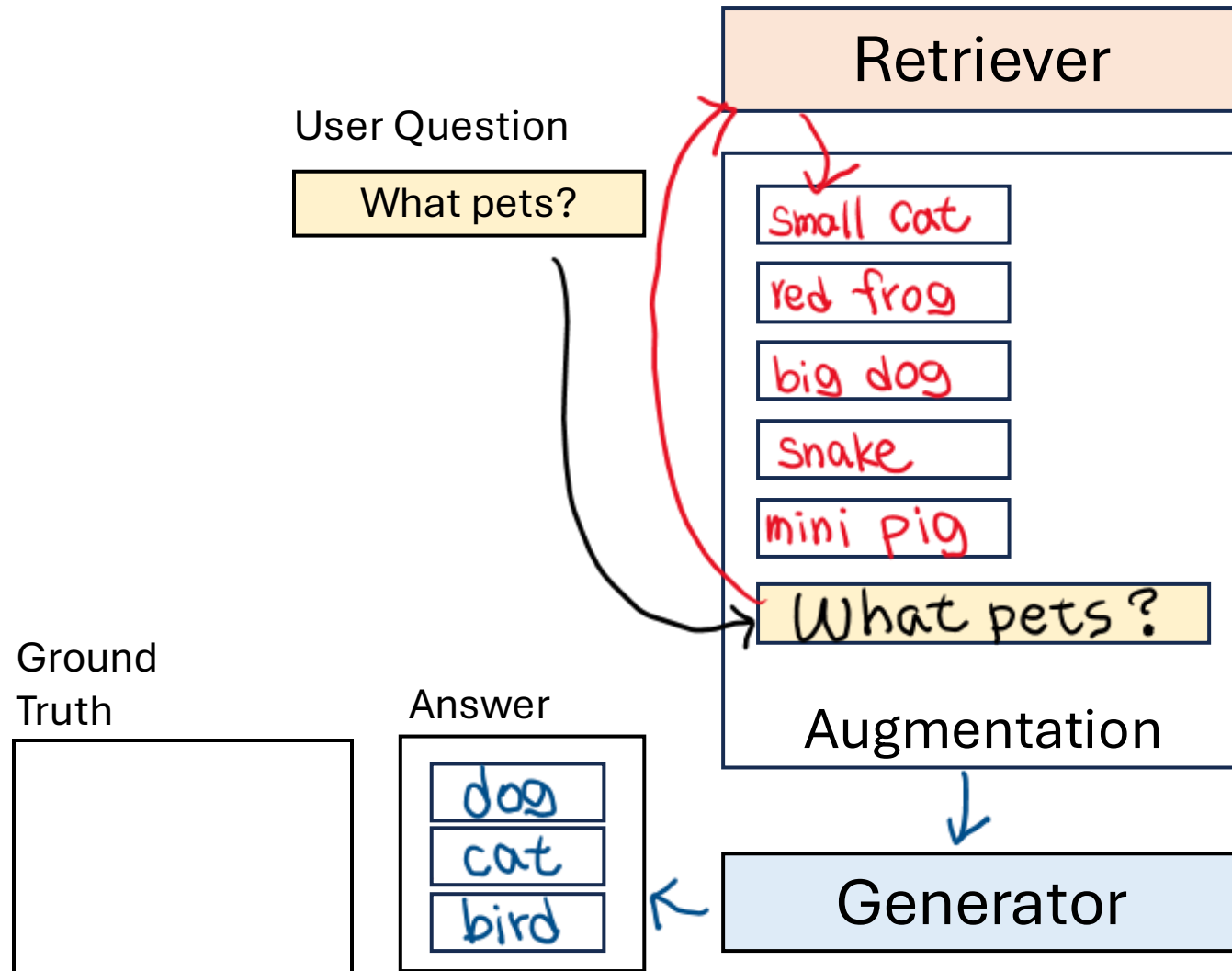
the signal to noise ratio of retrieved  
context

### context recall

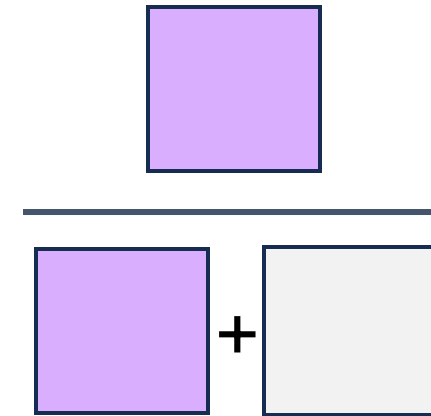
can it retrieve all the relevant information  
required to answer the question

# RAG

# Retriever Metrics



## Context Recall





# ragas score

## generation

### faithfulness

how factually accurate is  
the generated answer

### answer relevancy

how relevant is the generated  
answer to the question

## retrieval

### context precision

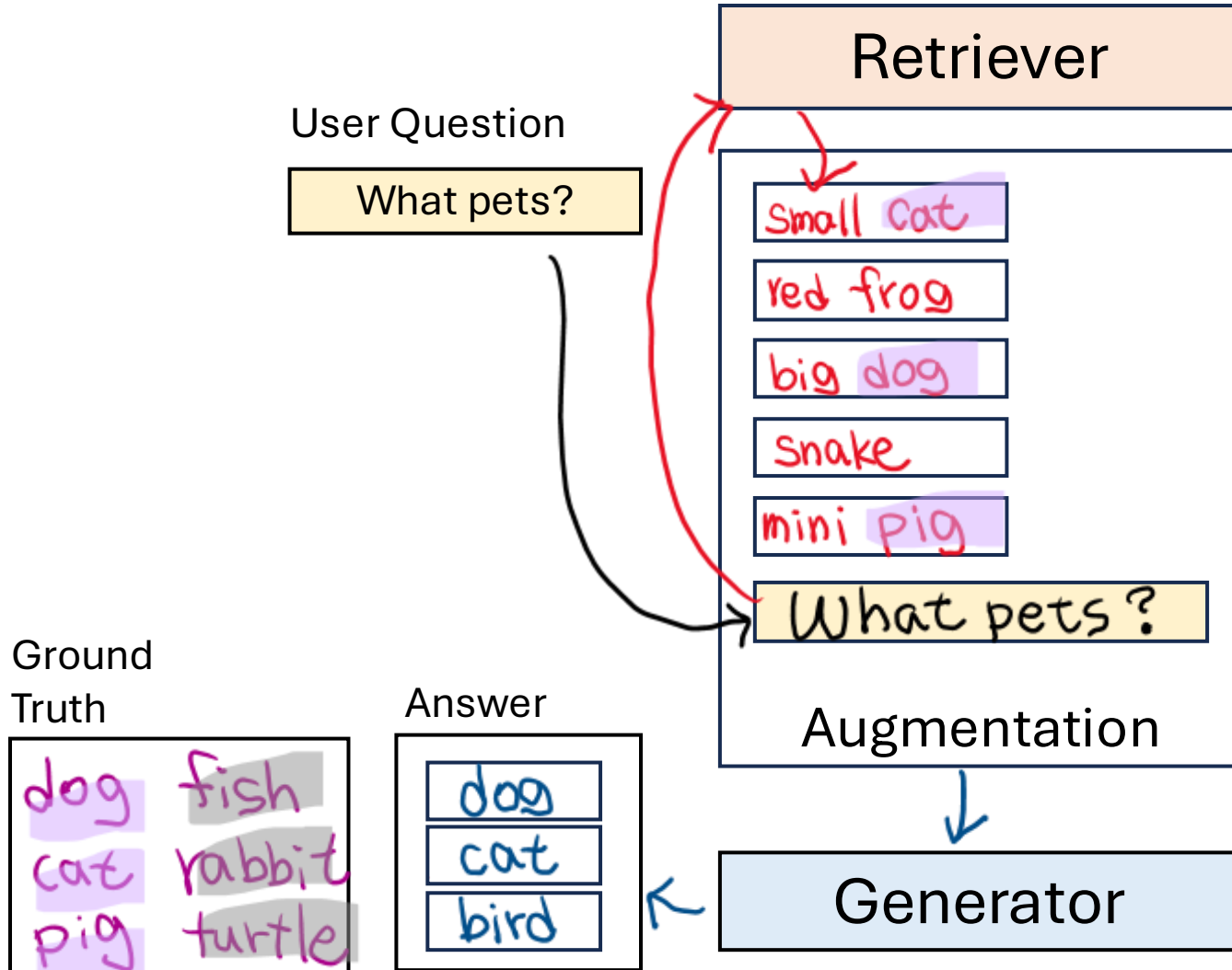
the signal to noise ratio of retrieved  
context

### context recall

can it retrieve all the relevant information  
required to answer the question

# RAG

# Retriever Metrics



## Context Precision

K	Relevant	TP@k	TP@k / k	
1				
2				
3				
4				
5				

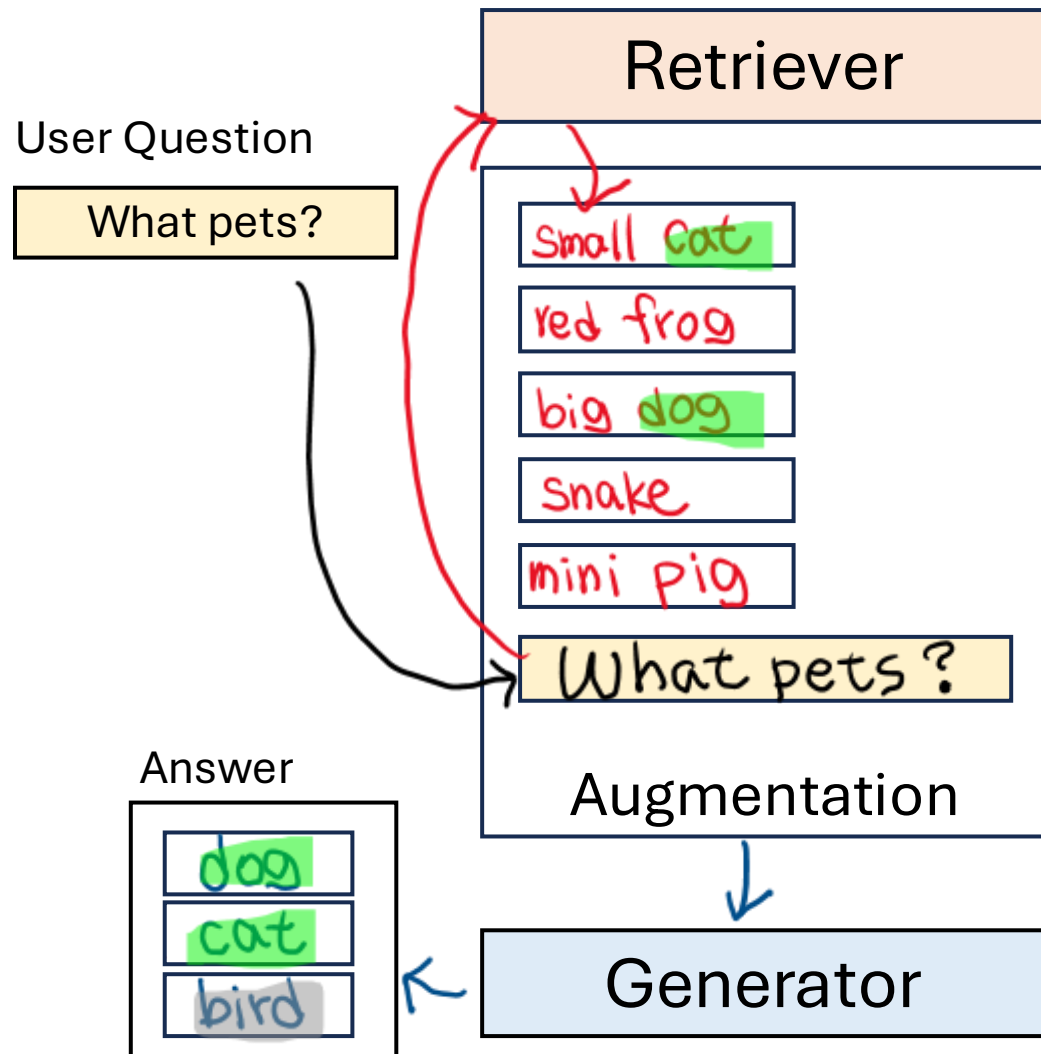
$$\sum \frac{1}{N}$$

# Precision

Given question, answer and context verify if the context was                      in arriving at the given answer. Give verdict as "1" if useful and "0" if not with json output.

Cost

# Where are LLMs used for Generator metrics?

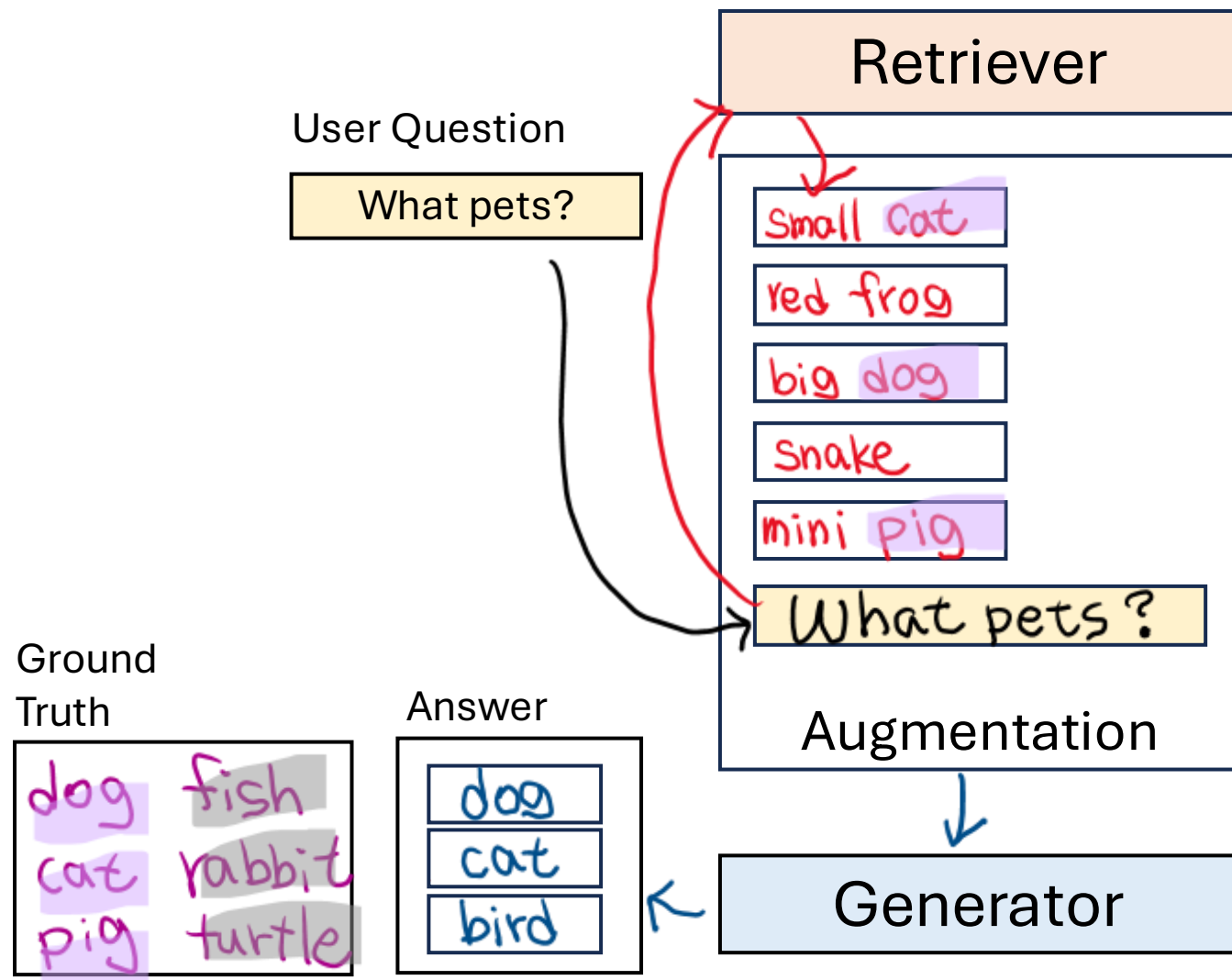


Retrieval

Generation

Evaluation

# Where are LLMs used for Retriever metrics?



Retrieval

Generation

Evaluation

# Many More Metrics

- Context Utilization
  - Context Entities Recall
  - Noise Sensitivity
  - Answer Semantic Similarity
  - Answer Correctness
  - Aspect Critique
  - Domain Specific Evaluation
  - Summarization Score
- ... etc.

	Target?	LLM Eval?	Ground Truth?	Range?	Best?
Faithfulness					
Answer Relevancy					
Context Recall					
Context Precision					
Context Utilization					
Context Entities Recall					
Noise Sensitivity					
Answer Semantic Similarity					
Answer Correctness					
Aspect Critique					
Domain Specific Evaluation					
Summarization Score					



Q/A