mollie

How to Keep Your LLM Chatbots Real

A Metrics Survival Guide

Maria Bader

PyData Amsterdam 2025

About me

Maria Bader, Ph.D.

Senior Data Scientist @ Mollie

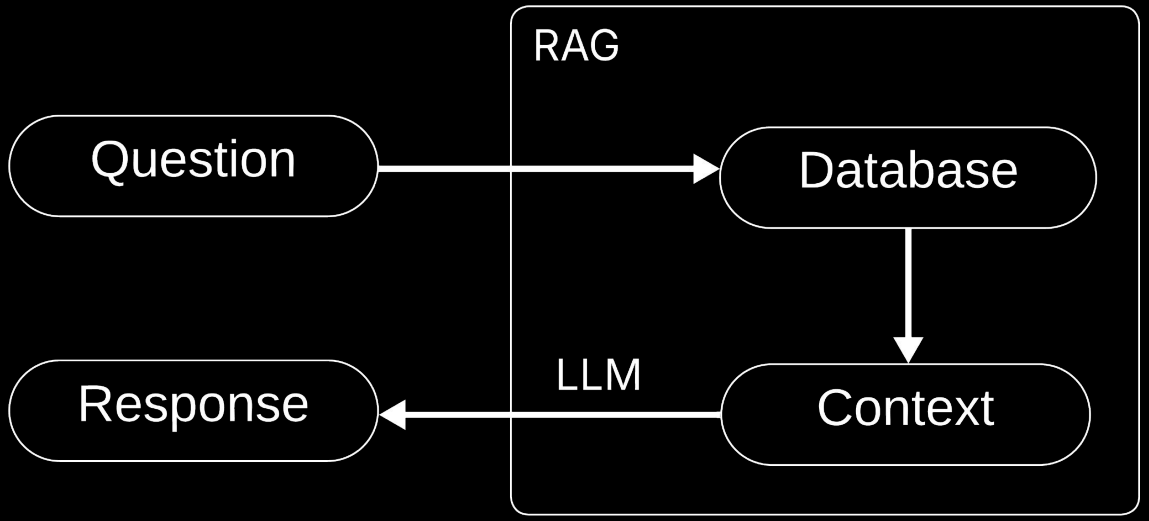
Formerly obsessed with tabular data.
Now full-time LLM whisperer.

RAG, Agentic Chatbots, function calling, Audio & Video analysis, OCR, MCP, ...

Then ...

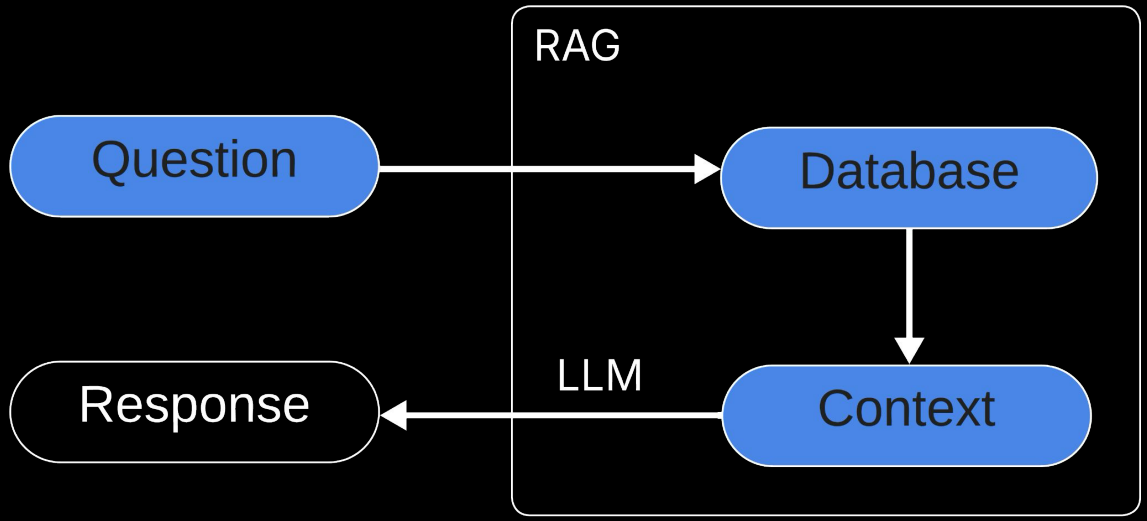
… now

Retriever Augmented Generation



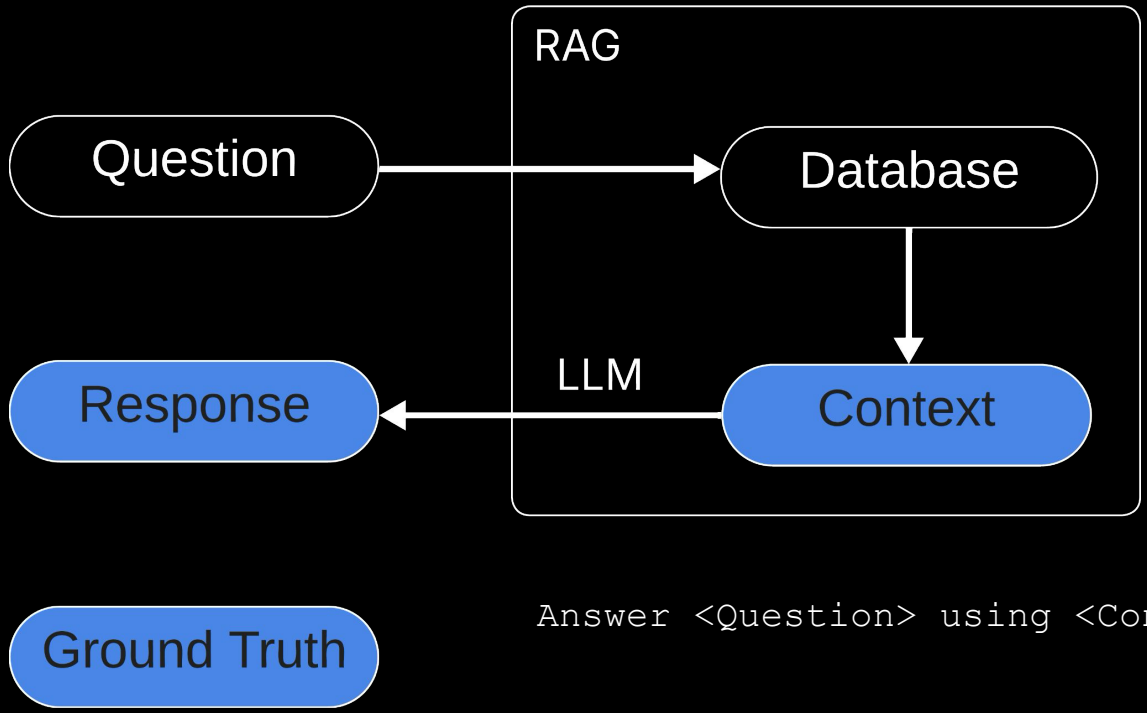

```
Answer <Question> using <Context>
```

Evaluation flow: Retrieval

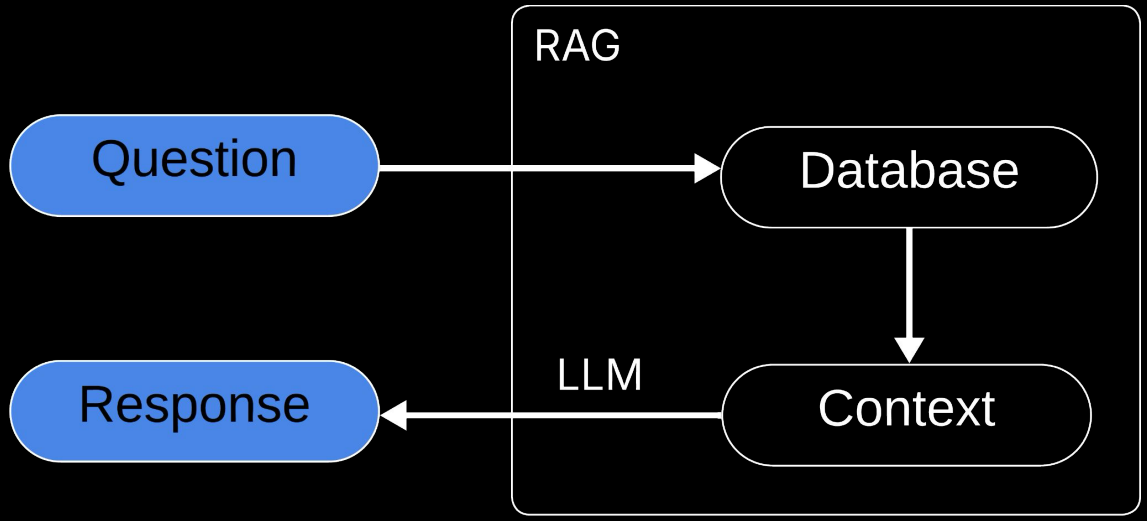



```
Answer <Question> using <Context>
```

Evaluation flow: Generation



Evaluation flow: End-to-end




```
Answer <Question> using <Context>
```

Our toolbox for chatbot evaluation

LLM-as-a-judge prompts

Advanced RAG evaluation

by ExplodingGradients

Other open source packages

Package	Synthetic Data	LLM-Judge	modular	RAG	Agentic	Bench-marks	General purpose
Ragas	single & multi	✅	✅	✅	✅	✅	✅
DeepEval	single & multi	✅	✅	✅	✅	✅	✅
LangChain	single	Ragas wrapper		✅	✅	✅	✅
MLFlow	❌	✅	❌	✅	✅	✅	✅
Haystack	❌	✅	✅	✅	❌	✅	❌
Llama Index	single & multi	✅	✅	✅	✅	❌	❌

... and many more

Introducing: CatBot

WIKIPEDIA
Simple English

Cat

272 languages

Page Talk Read Change Change source View history Tools

From Simple English Wikipedia, the free encyclopedia

Cats, also called **domestic cats** (*Felis catus*), are small, carnivorous (meat eating) mammals, of the family Felidae.[3][4][5] Ca… domesticated (tamed) for nea…

Cat (*Felis catus*)

Overview [change | cha…

Domestic cats may be cal…

Behaviour [change | ch…

Cats are active carnivores… mostly hunt small mamm… a wide variety of anima… main method of hunting…

History [change | change source]

In the past, mostly in Egypt, people kept cats because the cats hunted and ate mice and rats. The oldest evidence of cats kept as pets is from the Mediterranean island of Cyprus, around 7500 BC. Ancient Egyptians worshipped cats as gods, and often mummified them so they could be with their owners "for all of eternity".[13]

Today, people often keep cats as pets. There are also domestic cats which live without being cared for by people. These cats are called "feral cats" or "stray cats". The cats started becoming pets when the ancient Egyptians were around.

Creating a synthetic test set

Human generated

AI generated

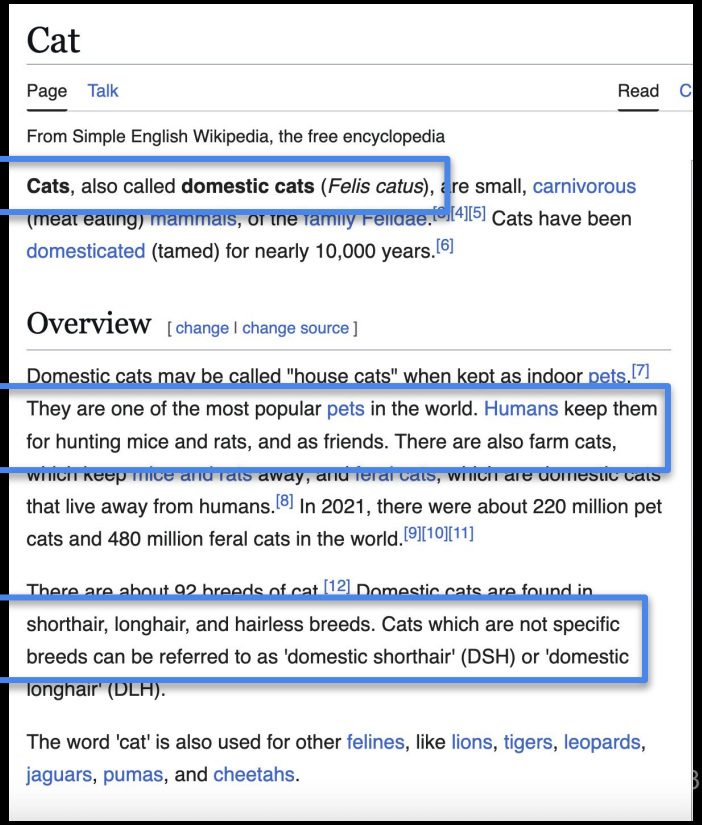
Single-context generation

"Based on <context>, generate a question"

What is the Latin name for cat?

What are the common purposes humans keep domestic cats for?

What is a domestic shorthair cat?



Cat

Page Talk Read C

From Simple English Wikipedia, the free encyclopedia

Cats, also called **domestic cats** (*Felis catus*), are small, carnivorous (meat eating) mammals, of the family Felidae.[3][4][5] Cats have been domesticated (tamed) for nearly 10,000 years.[6]

Overview [change | change source]

Domestic cats may be called "house cats" when kept as indoor pets.[7] They are one of the most popular pets in the world. Humans keep them for hunting mice and rats, and as friends. There are also farm cats, which keep mice and rats away, and feral cats, which are domestic cats that live away from humans.[8] In 2021, there were about 220 million pet cats and 480 million feral cats in the world.[9][10][11]

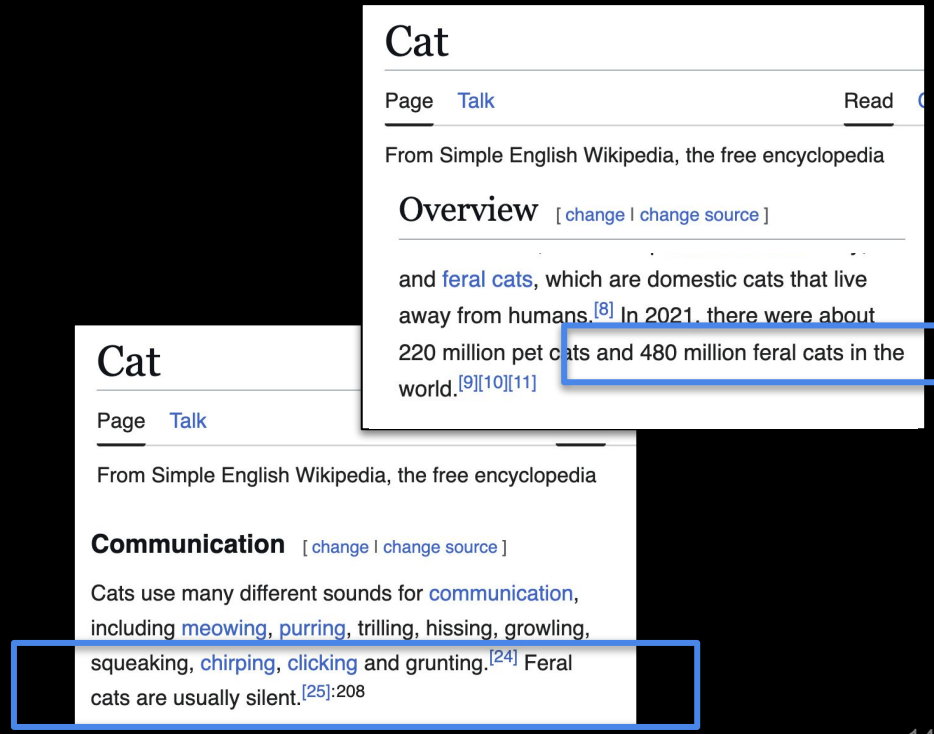
There are about 92 breeds of cat.[12] Domestic cats are found in shorthair, longhair, and hairless breeds. Cats which are not specific breeds can be referred to as 'domestic shorthair' (DSH) or 'domestic longhair' (DLH).

The word 'cat' is also used for other felines, like lions, tigers, leopards, jaguars, pumas, and cheetahs.

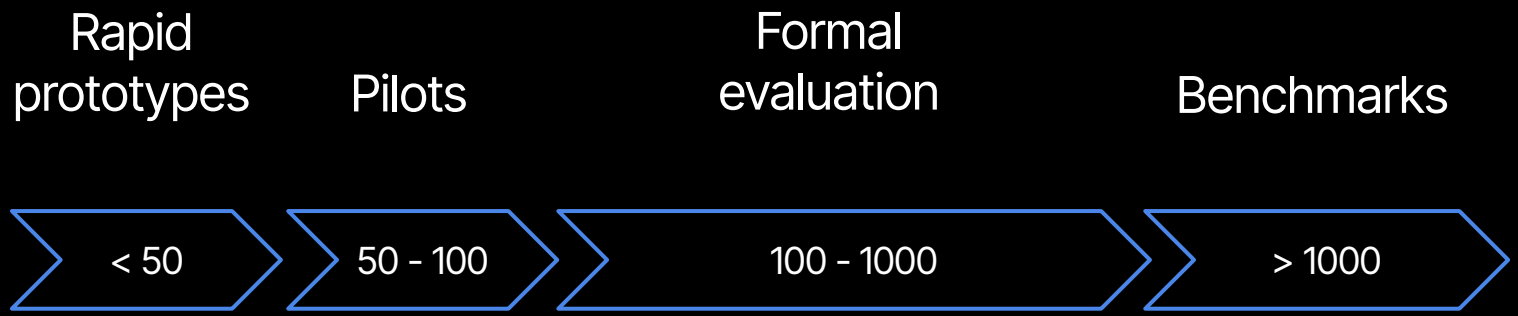
Multi-context generation

.... uses knowledge graph for context aggregation

How many feral cats are there and
are they as vocal as pet cats?



Choosing Your Test Set Size



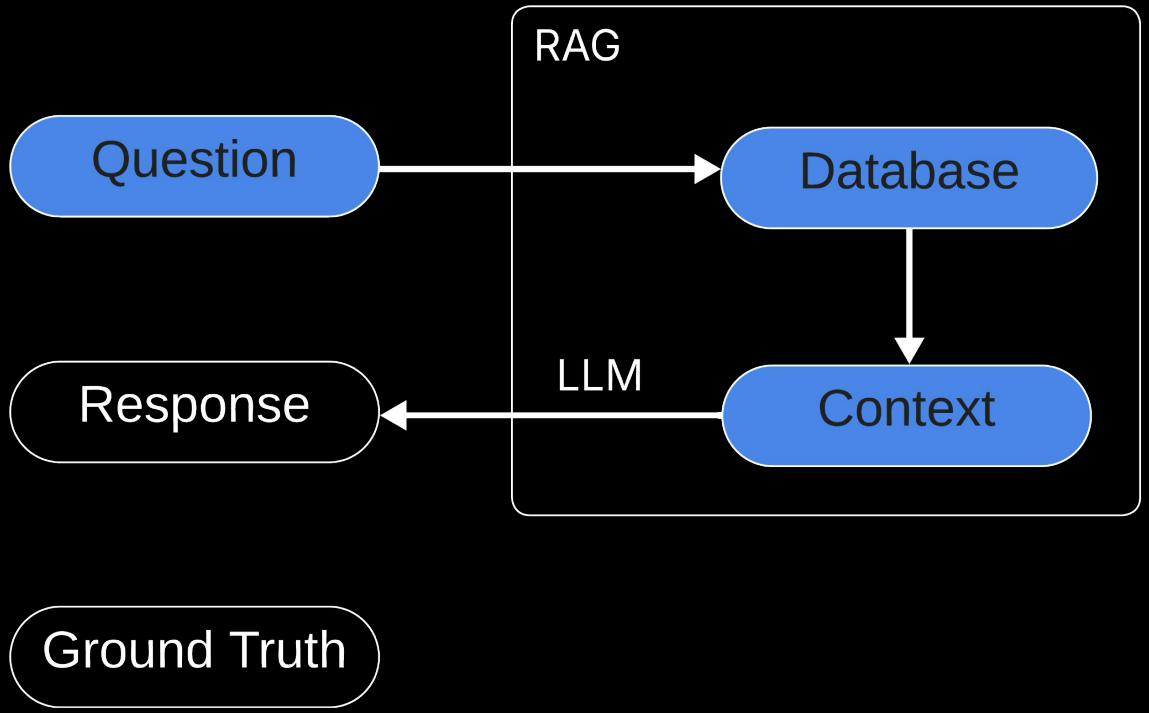
The evaluation flow

Synthetic	From RAG	
Question	Response	Contexts
What is the	A domestic ..	A
What age are....	The age where..	B
How do the ...	Historically,..	C
...	...	...

Score
1
0.5
0.8
...

(Weighted) Average Score
0.8

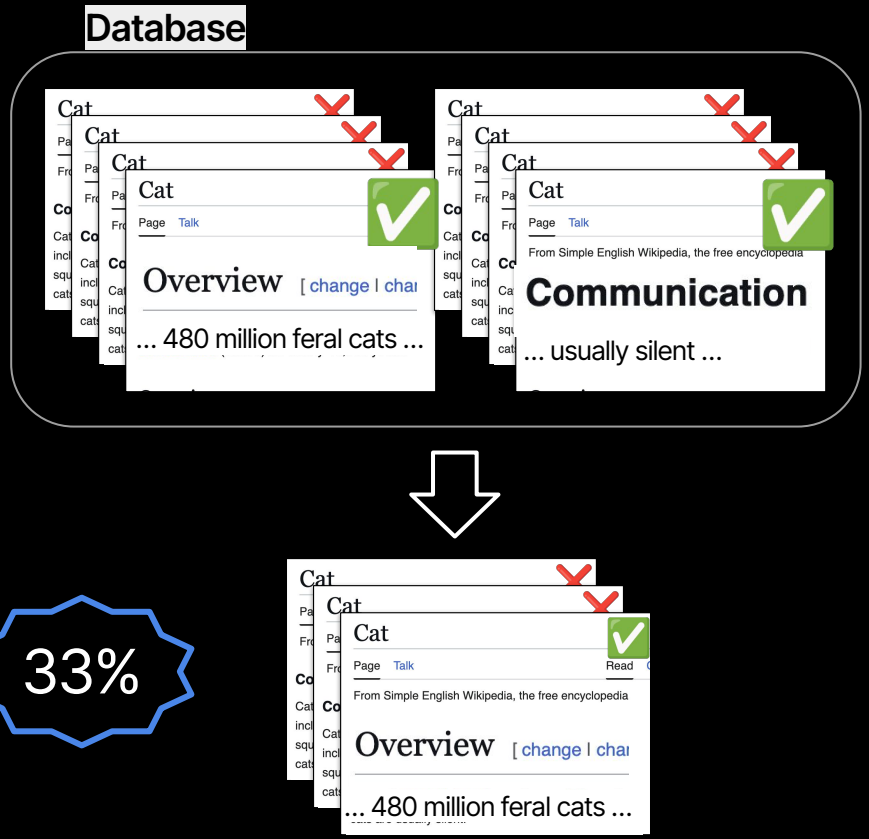
Retrieval: Context Precision and Recall



Context Precision

Proportion of relevant contexts
among all retrieved context.

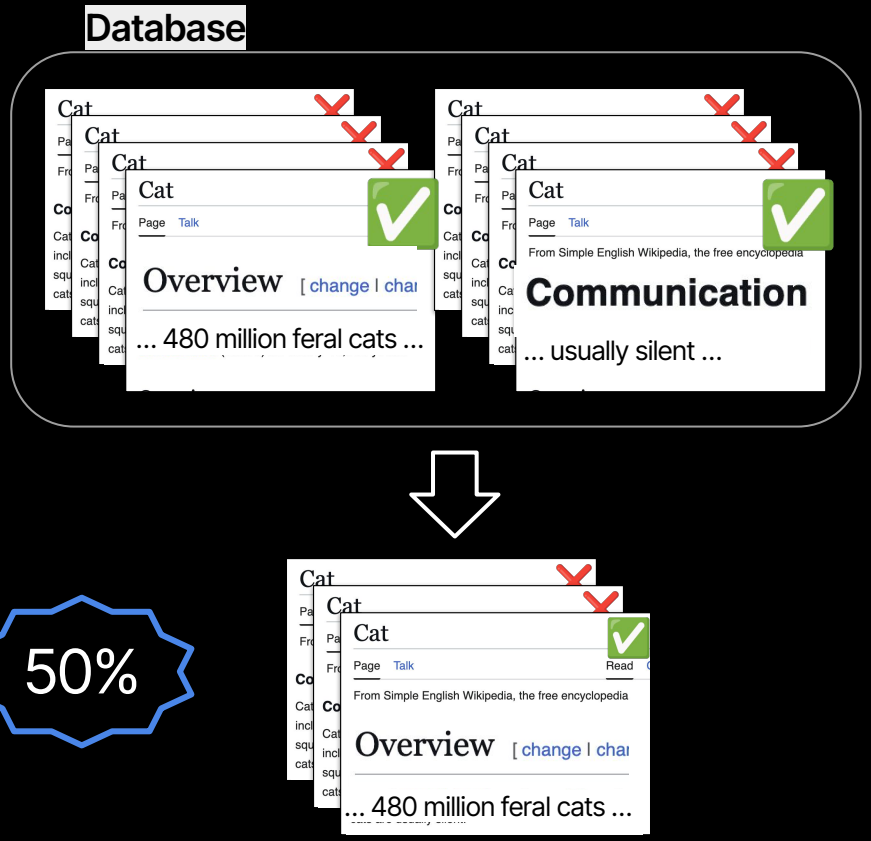
How many feral cats are there
and
are they as vocal as pet cats?



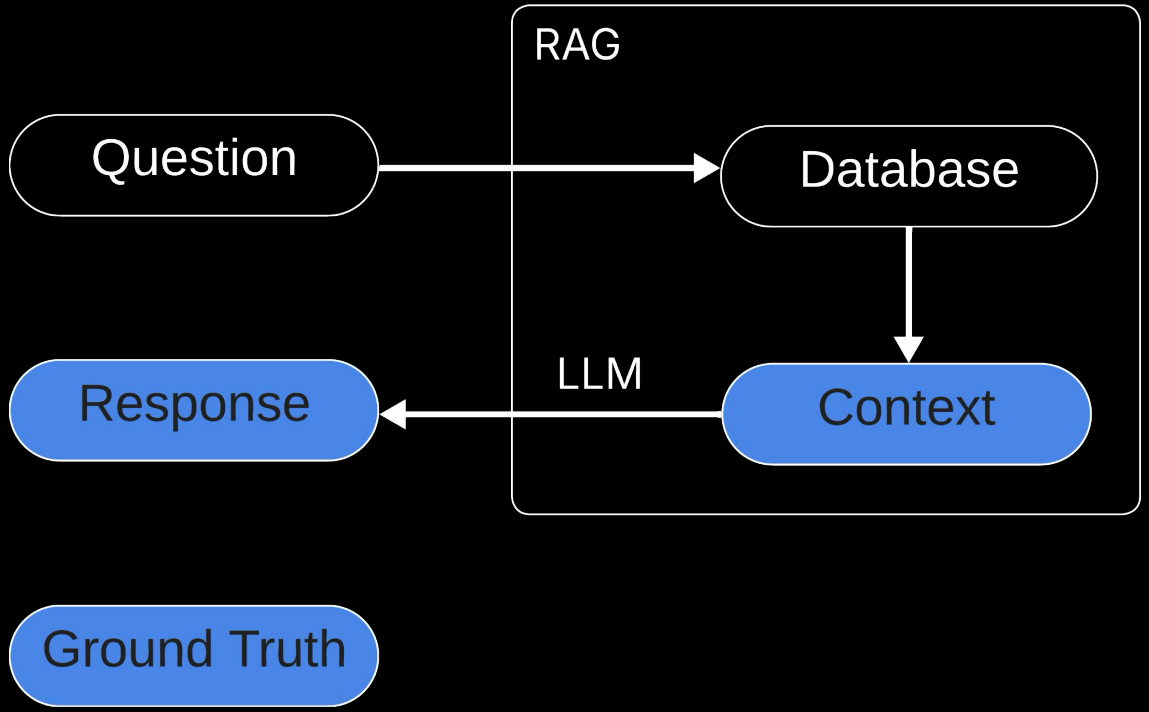
Context Recall

Proportion of successfully retrieved relevant contexts among all retrieved context.

How many feral cats are there and
are they as vocal as pet cats?



Generation: Faithfulness and Correctness



Faithfulness vs Correctness

What sounds do cats make?

Context:

Cat

Page Talk

Cats bark

Ground Truth:

Cats meow

Context Documents

Faithfulness

Correctness

Response

Ground Truth

woufff

Context says: bark, but should meow

quaaack

meooow

Faithfulness

What is the Latin name for cat?

The Latin name for cat is *Felis catus*.

Context:

Cat

Page Talk

From Simple English Wikipedia, the free encyclopedia

Cats, also called **domestic cats** (*Felis catus*),

LLM-as-a-judge

100 %

```
Grade faithfulness in the  context:
- 0 %   | none of the claims are supported
- 25 %  | some of the claims are supported
- 50 %  | half of the claims are supported
- 75 %  | most of the claims are supported
- 100 % | all of the claims are supported
```

ragas

100 %

```
Extract statements
```

```
Check if represented in context
```

Average

✅ **Statement 1:** The Latin name ...

Faithfulness

What is the Latin name for cat?

The Latin name for cat is *Felis catus*.
The Latin plural of *felis* is *feles*.

Context:

Cat

Page Talk

From Simple English Wikipedia, the free encyclopedia

Cats, also called **domestic cats** (*Felis catus*),

LLM-as-a-judge

50 %

```
Grade faithfulness based on  context:
- 0 %   | none of the claims are supported
- 25 %  | some of the claims are supported
- 50 %  | half of the claims are supported
- 75 %  | most of the claims are supported
- 100 % | all of the claims are supported
```

ragas

50 %

```
Extract statements
```

```
Check if represented in context (bool)
```

Average

✅ **Statement 1**: The Latin name ...
❌ **Statement 2**: The Latin plural ...

Correctness

What is the Latin name for cat?

The Latin name for cat is *Felis catus*.
The Latin plural of *felis* is *feles*.

Ground Truth:

The scientific name for the domestic cat is Felis catus. Additionally, in Latin, the plural form of the word felis is feles.

LLM-as-a-judge

100 %

```
Grade correctness based on  ground truth:
- 0 %   | none of the claims are supported
- 25 %  | some of the claims are supported
- 50 %  | half of the claims are supported
- 75 %  | most of the claims are supported
- 100 % | all of the claims are supported
```

ragas

Extract statements

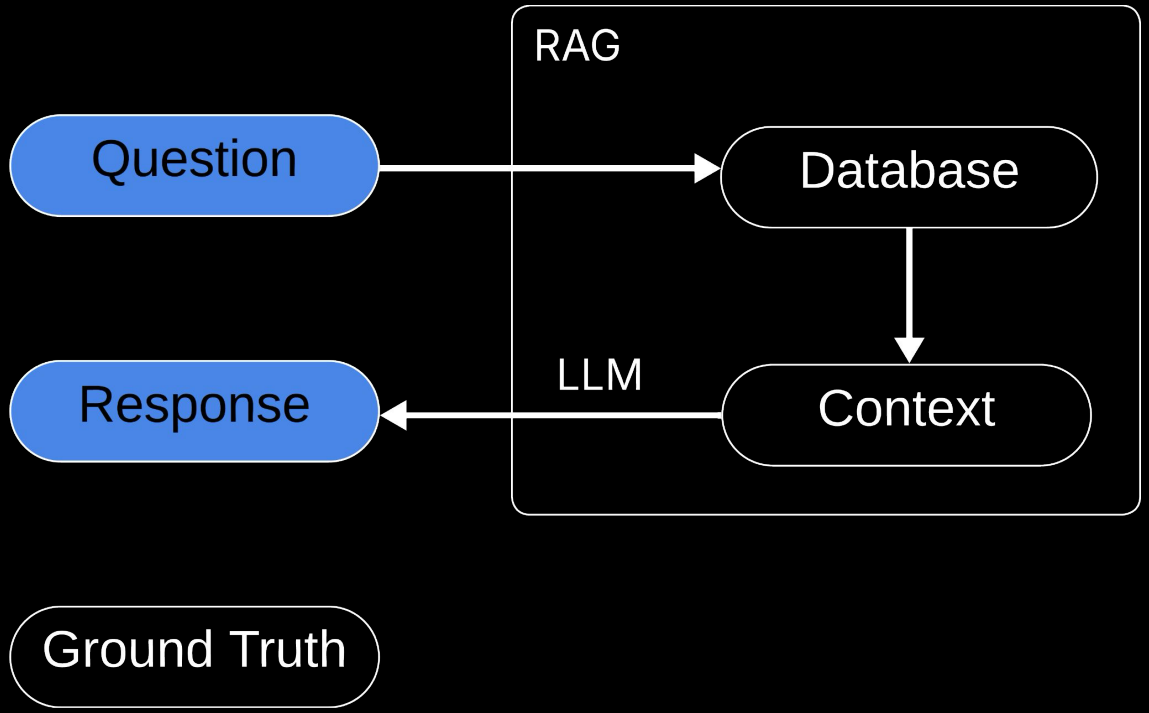
Check if represented in ground truth (TP, FP, FN)

Factual similarity (F1) 100 %

Semantic similarity 80 %

90 %

Evaluation flow: End-to-end



Answer Relevance

What is the Latin name for cat?

The Latin name for cat is *Felis catus*.
The Latin plural of *felis* is *feles*.

`Generate questions from response`

`Cosine similarity to user question`

Average

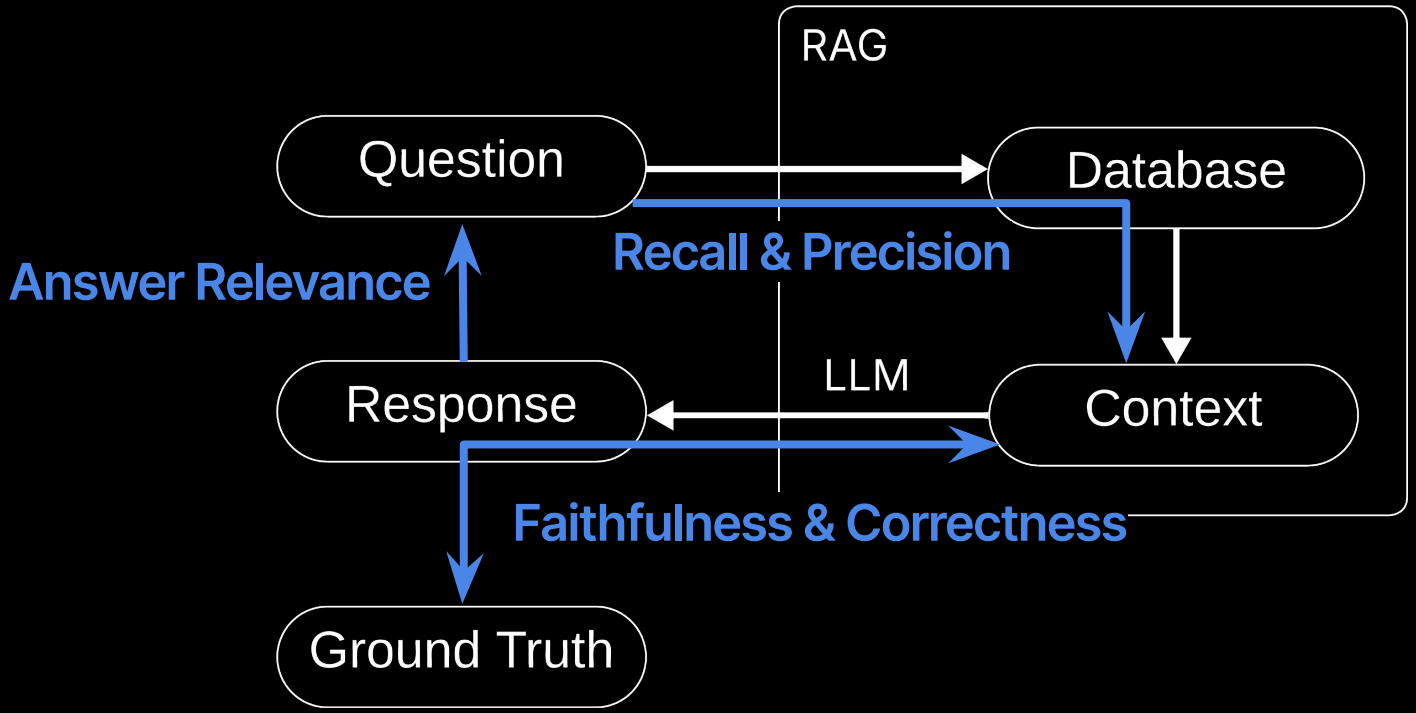
88 %

Question 1: What is the Latin name for a cat? 100 %

Question 2: What does "Felis catus" mean? 85 %

Question 3: What is the plural of "felis"? 80 %

Full circle moment



Beyond RAG evaluation

Advanced RAG

HHem
Faithfulness with Vectra's open classifier model

Multimodal RAG metrics

Role adherence

General purpose

Semantic similarity

Summarization Score
Multi-prompt based on keyphrases

Traditional NLP metrics
BLEU, ROUGE, ...

Agent/MCP

Tool call Accuracy
How well is the correct tool called?

Argument correctness
Are the correct arguments used?

Topic Adherence
How well does the agent stay on topic?

Agent Goal accuracy
How well is the user's goal achieved?

Thank you for your attention

Let’s connect

Github: Catbot

Maria Bader, Ph.D.
Senior Data Scientist at Mollie B.V.