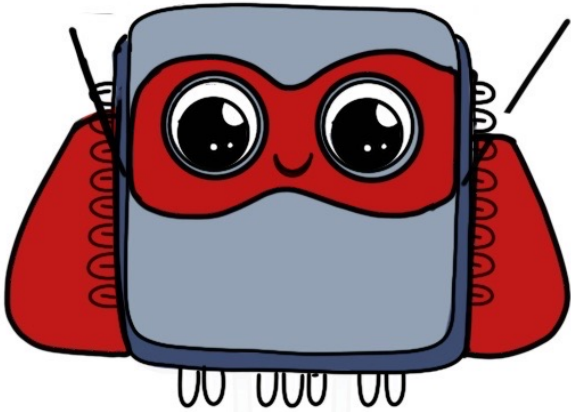


# RETRIEVAL AUGMENTED GENERATION

# RAG

THE  
SUPER INGREDIENT  
TO MAKE LARGE  
LANGUAGES MODELS  
MORE ROBUST AND  
RELIABLE



BY ORSI NAGY

As we start using generative AI at and outside work, there are a number of questions about their trustworthiness

I'D LIKE TO USE  
LARGE LANGUAGE  
MODELS, BUT IS THE  
OUTPUT CORRECT?

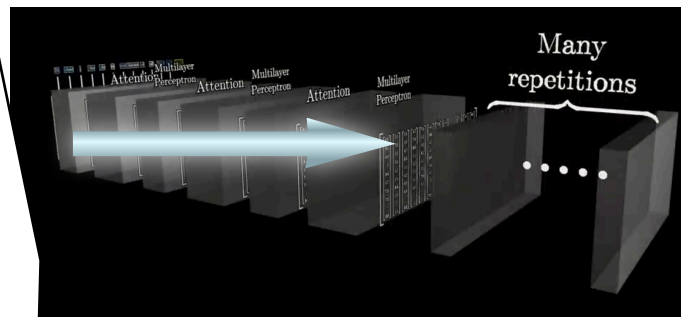
I WOULD LIKE  
THEM TO REFER TO  
INTERNAL, UP-TO-  
DATE DATA AND  
KNOWLEDGE

HOW CAN I  
KNOW IF AND  
WHEN IT IS  
CORRECT?

I'D LIKE TO  
VERIFY, TRACE,  
OR CITE  
SOURCES.



AS THE INFORMATION MOVES THROUGH  
THE MODEL, PARAMETERS ARE  
ACTIVATED, AND MANY MANY  
CALCULATIONS ARE LAUNCHED

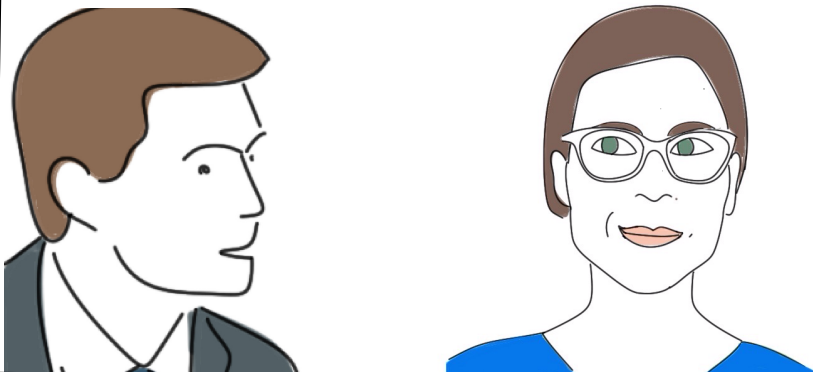


IT IS OFTEN NOT POSSIBLE TO  
TRACE ELEMENTS BACK TO THEIR  
ORIGINAL SOURCES, LEADING TO  
ISSUES WITH ACCOUNTABILITY AND  
TRUST

**THE  
WRITING AND REASONING  
CAPACITIES ARE IMPRESSIVE  
THOUGH**

THESE  
MODELS SOUND  
SO INTELLIGENT.  
CAN'T WE JUST  
TRUST THEM?

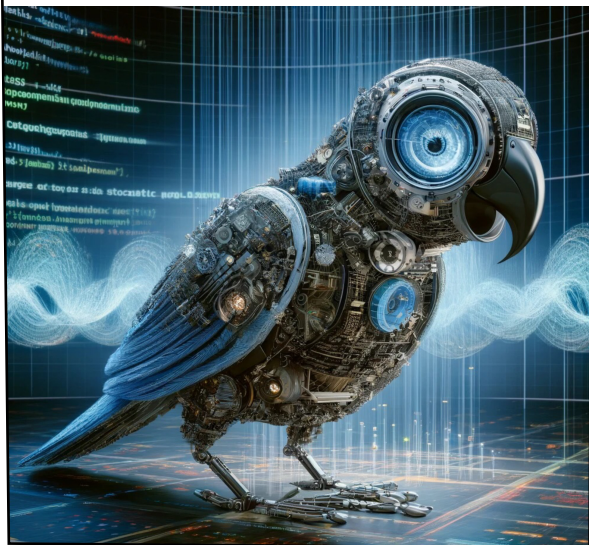
UNFORTUNATELY NO.  
THEY GENERATE TEXT  
BASED ON PROBABILITIES  
RATHER THAN TRUE  
UNDERSTANDING OR  
PRECISE REFERENCES.



THIS IS WHY SOMETIMES THEY ARE CALLED

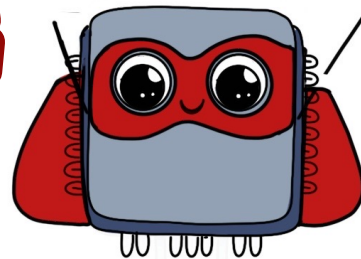
# stochastic parrots

- GENERATE OUTPUTS THAT SOUND LIKE,
- REPLICATE INFORMATION WITHOUT UNDERSTANDING,
- USING PROBABILITIES.



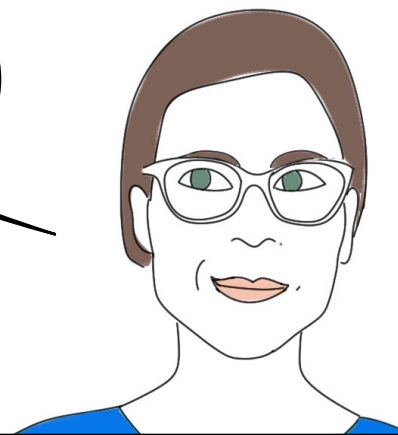
**THIS IS WHERE**

# RAG

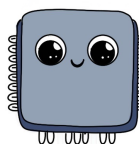
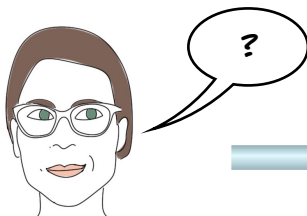


**COMES IN!**

WHEN WE USE RAG, WE ADD A RETRIEVAL (SEARCH) SYSTEM TO AN LLM THAT PULLS UP RELEVANT, RELIABLE INFORMATION



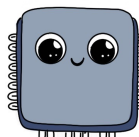
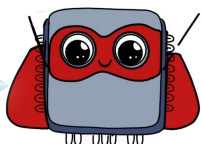
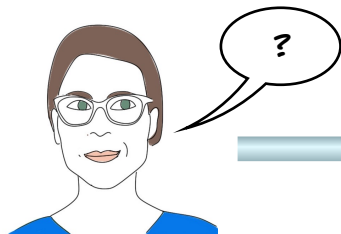
**LLM**



**GENERATED ANSWER**

Generation

LARGE LANGUAGE MODEL



**GENERATED ANSWER**

Generation

LARGE LANGUAGE MODEL

QUESTION



Specific (private) Knowledge Base

**LLM with**

# RAG

## IF you use Retrieval Augmented Generation...

Your prompt

**You can see immediately the source documents...**

Sources

Document 1. You can see the content used for generating the response.

Document 2. You can see the content used for generating the response.

Document 3. You can see the content used for generating the response.

View 2 more

by RAG Agent x source:    by RAG Agent x source:    by RAG Agent x source:    by RAG Agent x source:

Answer

>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

As well as the generated answer

## With just one click you can go back to the original document to verify the source

Sources

Document 1. You can see the content used for generating the response.

Document 2. You can see the content used for generating the response.

Document 3. You can see the content used for generating the response.

View 2 more

by RAG Agent x source:    by RAG Agent x source:    by RAG Agent x source:    by RAG Agent x source:

Answer

**IF you have well-organised documents with good metadata about the type of documents, date, security level, etc., your search will be even better!**

My Data

Name of dataset\*

Name of agent\*

Link to dataset\*

Description

Publisher\*

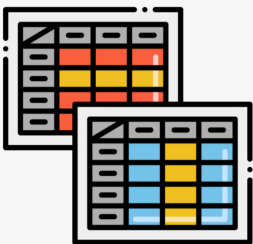
Data Theme

Access\*

## you can retrieve and use all sorts of data...



MANY MANY MANY DOCUMENTS



NUMBERS IN EXCEL FILES OR OTHER STRUCTURED DATASETS



OR EVEN KNOWLEDGE GRAPHS!

Knowledge graph

HOORAY!

THIS IS REASSURING

I LIKE THAT I CAN ACCESS MANY SOURCES AT ONCE

I ALREADY HAVE SOME IDEAS WHAT DATA I'D LIKE TO USE!