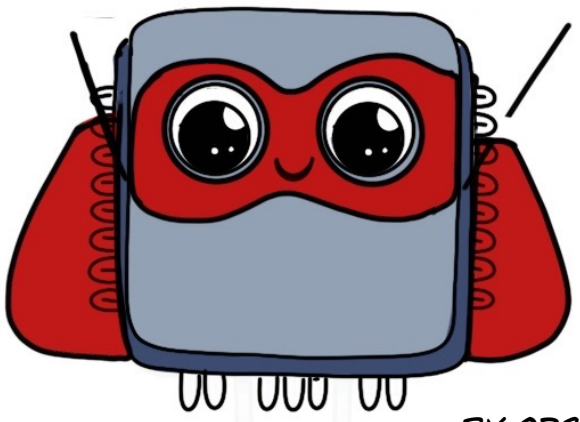
RETRIEVAL AUGMENTED GENERATION

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

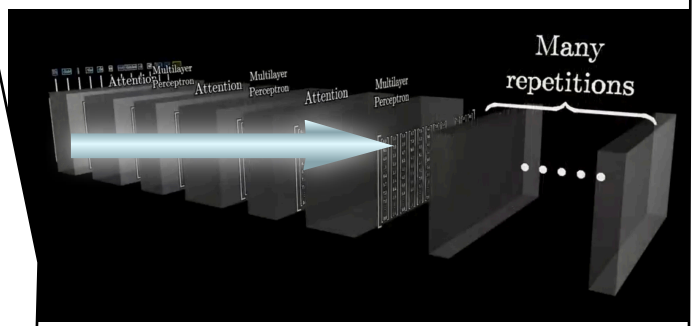
I'D LIKE TO USE LARGE LANGUAGE MODELS, BUT IS THE OUTPUT CORRECT?

HOW CAN I KNOW IF AND WHEN IT IS CORRECT?

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

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

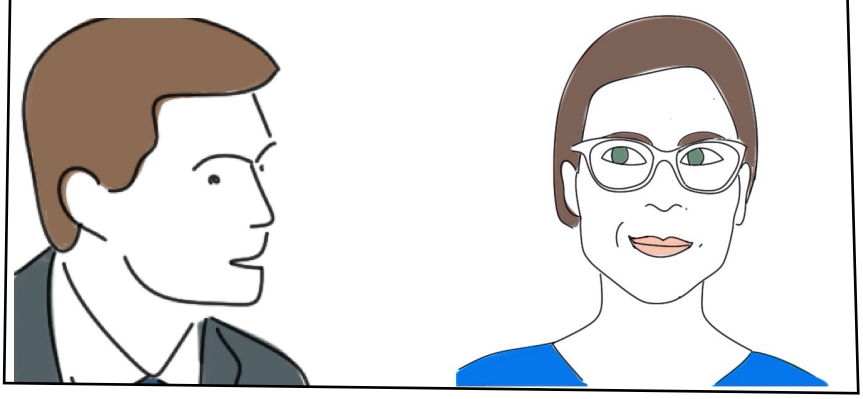
AS THE INFORMATION MOVES THROUGH THE MODEL, PARAMETERS ARE ACTIVATED, AND MANY MANY CALCULATIONS ARE LAUNCHED



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

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.



THE WRITING AND REASONING CAPACITIES ARE IMPRESSIVE THOUGH

THIS IS WHY SOMETIMES THEY ARE CALLED

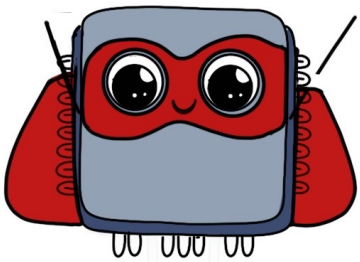
stochastic parrots

- GENERATE OUTPUTS THAT SOUND LIFELIKE,

- REPLICATE INFORMATION WITHOUT UNDERSTANDING,

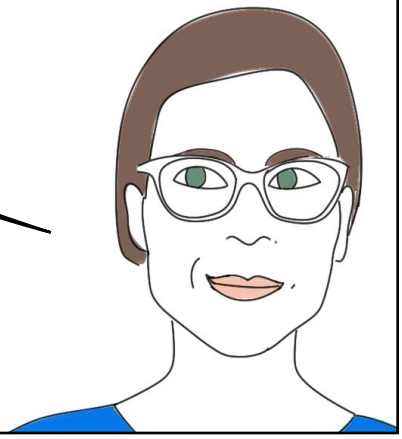
- USING PROBABILITIES.

THIS IS WHERE RAG

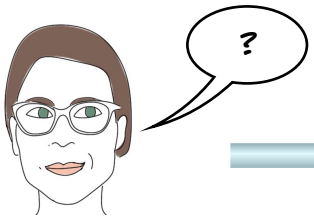


COMES IN!

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



LLM

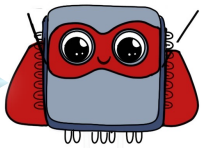


LARGE LANGUAGE MODEL

Generation

GENERATED ANSWER

QUESTION



INTELLIGENT SEARCH

Retrieval

Specific (private) Knowledge Base

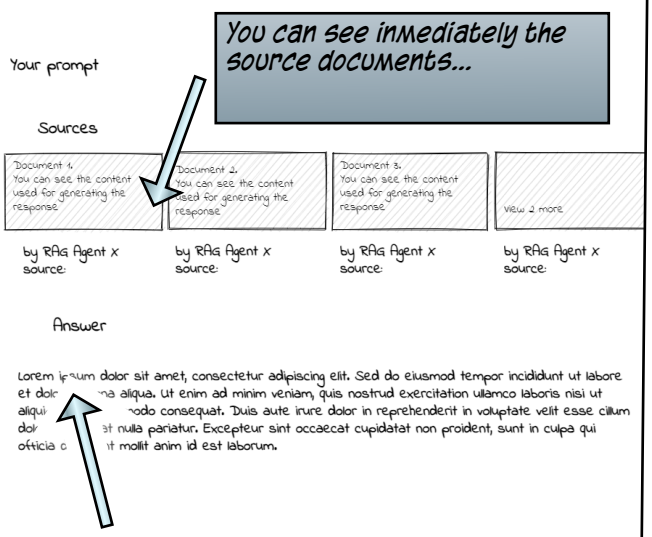
LARGE LANGUAGE MODEL

Generation

GENERATED ANSWER

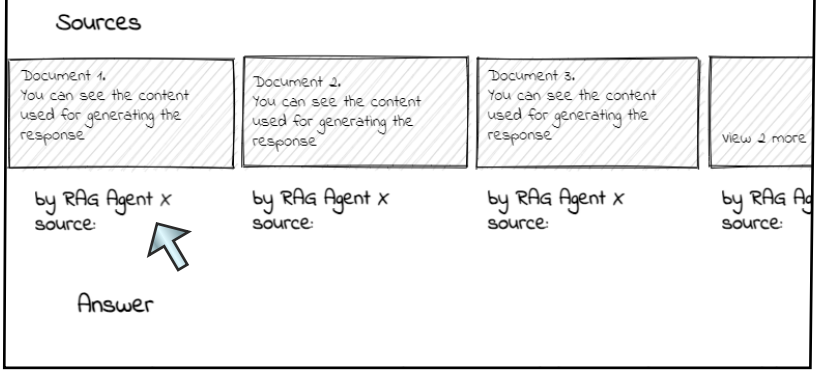
LLM with RAG

If you use Retrieval Augmented Generation...

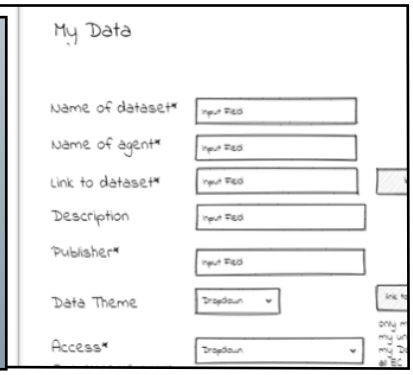


As well as the generated answer

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

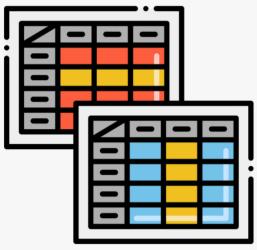


If you have well-organised documents with good metadata about the type of documents, date, security level, etc., your search will be even better!



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!