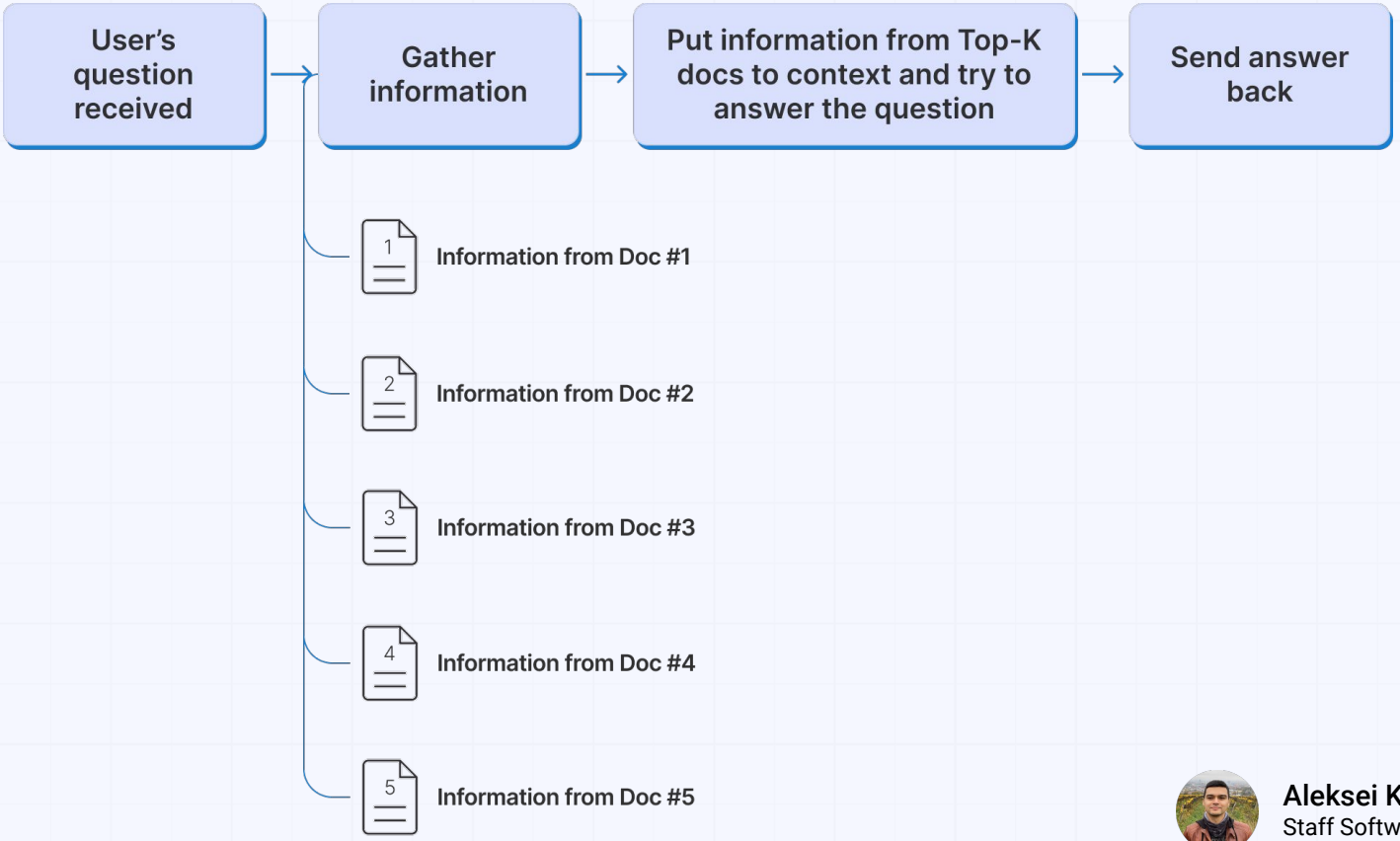
Modern Paradigms and approaches in GenAI

Aleksei Kolesnikov
Staff Software Engineer

RAG paradigm using Top-K wording in simple words



Aleksei Kolesnikov
Staff Software Engineer

Common issues with Top-K

→ What if information is not relevant or outdated?

→ What if important information presented in other documents missed?

→ What if information contradicts other information?

It may fail in other scenarios as well!

Aleksei Kolesnikov
Staff Software Engineer

Re-ranking in RAG paradigm

Idea

Re-ranking is a simple but effective concept in RAG technology. First, you retrieve many documents (like 10). Then, a reranker model picks the top few (like 5) to use as language model context. This approach makes sense because the model getting the top contexts isn't trained for knowledge retrieval, so it's useful to have a smaller reranker model for specific RAG situations.

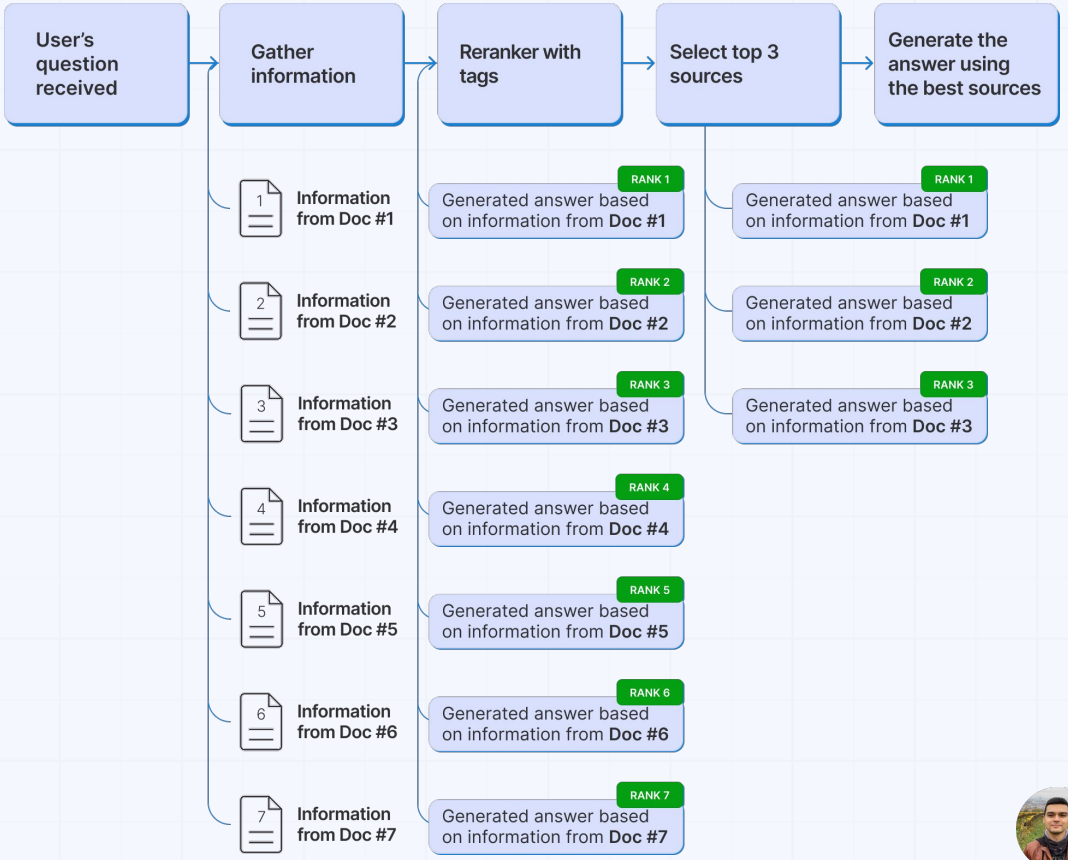
Re-ranking in RAG Advantages

1. Simple and powerful idea.

2. Provides better results than RAG in case you need up to date information.

Aleksei Kolesnikov
Staff Software Engineer

Re-ranking in RAG



Aleksei Kolesnikov
Staff Software Engineer

Self-reflective RAG or Self-RAG Paradigm

Idea

Criticise the information sources and information generated based on them, label the retrieval in several dimensions like [Relevant] - [Irrelevant] or [Contradicts] - [Not contradicts], [Answer] - [Utility]

Self-RAG operates through a sequential process

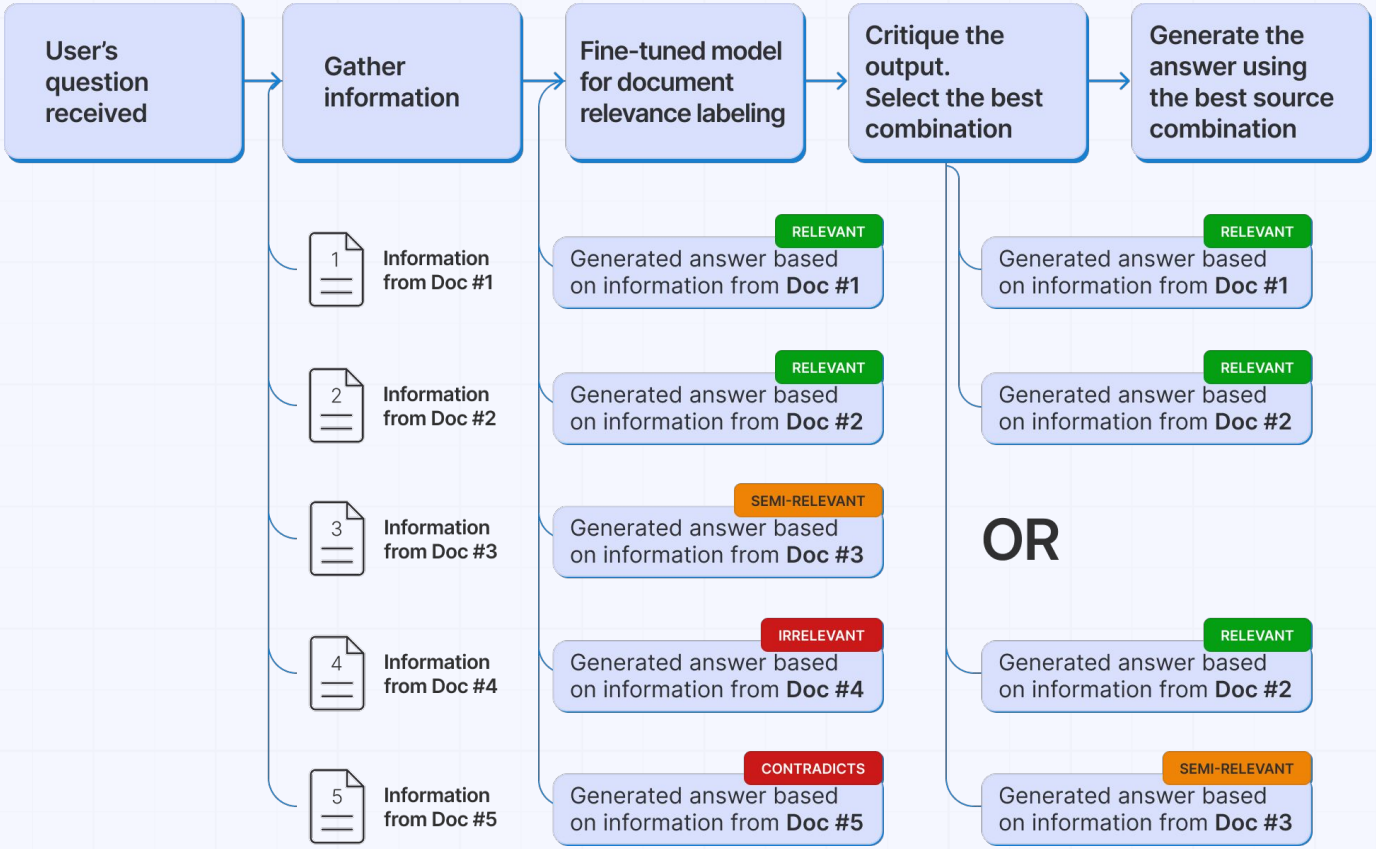
1. The method starts with training a basic language model (LM) to classify generated outputs.

2. Followed by this, the process facilitates the creation of continuations and special tokens.

*for more information you can visit: https://selfrag.github.io/

Aleksei Kolesnikov
Staff Software Engineer

Self-RAG


User's question received
Gather information
Fine-tuned model for document relevance labeling
Critique the output. Select the best combination
Generate the answer using the best source combination
1
Information from Doc #1
2
Information from Doc #2
3
Information from Doc #3
4
Information from Doc #4
5
Information from Doc #5
RELEVANT
Generated answer based on information from Doc #1
RELEVANT
Generated answer based on information from Doc #2
SEMI-RELEVANT
Generated answer based on information from Doc #3
IRRELEVANT
Generated answer based on information from Doc #4
CONTRADICTS
Generated answer based on information from Doc #5
RELEVANT
Generated answer based on information from Doc #1
RELEVANT
Generated answer based on information from Doc #2
OR
RELEVANT
Generated answer based on information from Doc #2
SEMI-RELEVANT
Generated answer based on information from Doc #3

Self-RAG advantages

- **Adaptive Passage Retrieval:** It ensures all relevant context is found within a set context window.
- **Improved Relevance:** It often outperforms embedding models in retrieving pertinent context.
- **Special token use:** It utilizes a special token system to aid relevance.
- **Superior performance:** It often surpasses similar models, even surprisingly outperforming ChatGPT in various tasks, which indicates potential for application with proprietary data.
- **Preservation of underlying Language Model:** Unlike methods like fine-tuning and **RLHF**, which can bias models, **Self-RAG** simply adds special tokens, leaving the original text generation unchanged.

Forward-Looking Active Retrieval Augmented Generation (FLARE)

Idea

Set up a process where you divide the information into separate sections and address each part individually. Also, ensure that you are using a current source of information, such as the internet.

Forward-Looking Active Retrieval Augmented Generation (FLARE) Advantages

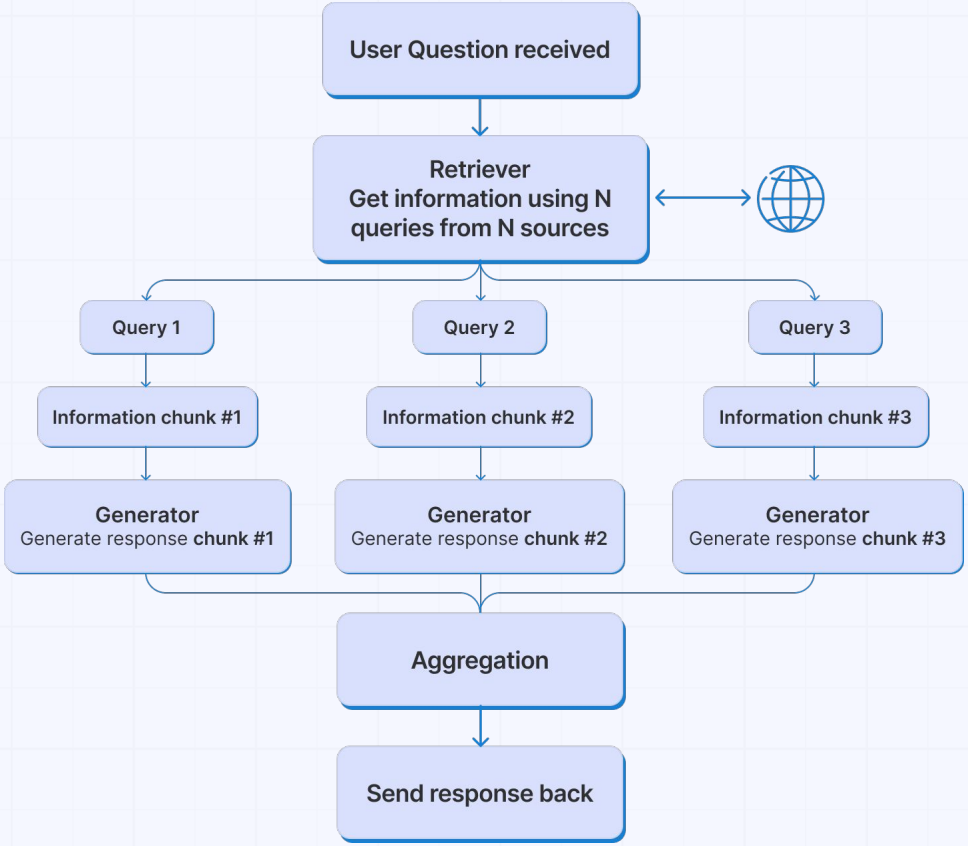
→ Works good if you need a summary

→ Can aggregate information from several sources

→ Can answer on questions to predict next steps.

Aleksei Kolesnikov
Staff Software Engineer

Forward-Looking Active Retrieval Augmented Generation (FLARE)

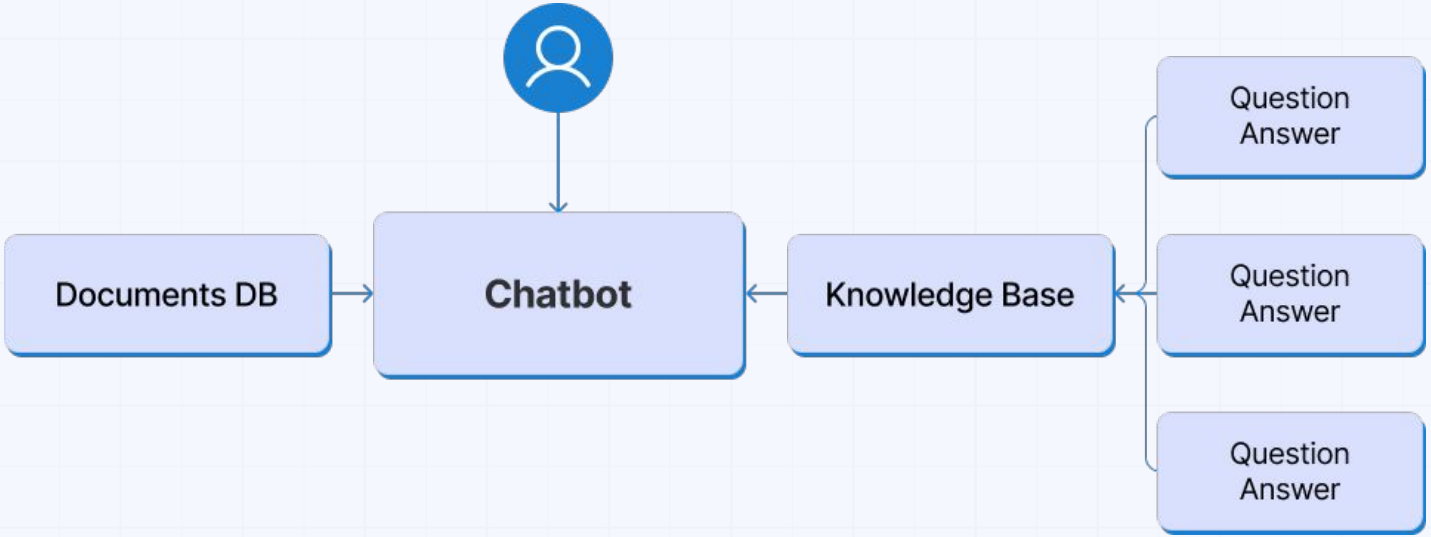


Aleksei Kolesnikov
Staff Software Engineer

Predefined Q&A. Knowledge Base

Aleksei Kolesnikov
Staff Software Engineer

Q&A Knowledge Database



Aleksei Kolesnikov
Staff Software Engineer

Dealingwith opinionated requests and information inputs

Aleksei Kolesnikov
Staff Software Engineer

Uncaptured Moments with the RAG Family

Idea

In my earlier post on the RAG family, accessible here I purposefully left out a potential problem with context.

What happens if the question or context is biased?
What if the information sources reiterate misleading data?

Under such circumstances, the performance of the RAG family might falter.

So, what's our workaround?

The answer is simple:

1. Filter the information!
2. Rate the response!

Aleksei Kolesnikov
Staff Software Engineer

System 2 Attention (S2A) by Meta

Idea

Use specially tuned language models to rewrite the context. S2A does this mainly by eliminating irrelevant text.

As a result, it allows the language models to make precise decisions about what parts of the input to home in on before churning out a response.

Why

The built-in attention mechanism isn't perfect and can sometimes pull in needless info into the context. This gets particularly tricky when a specific entity pops up several times in the context.

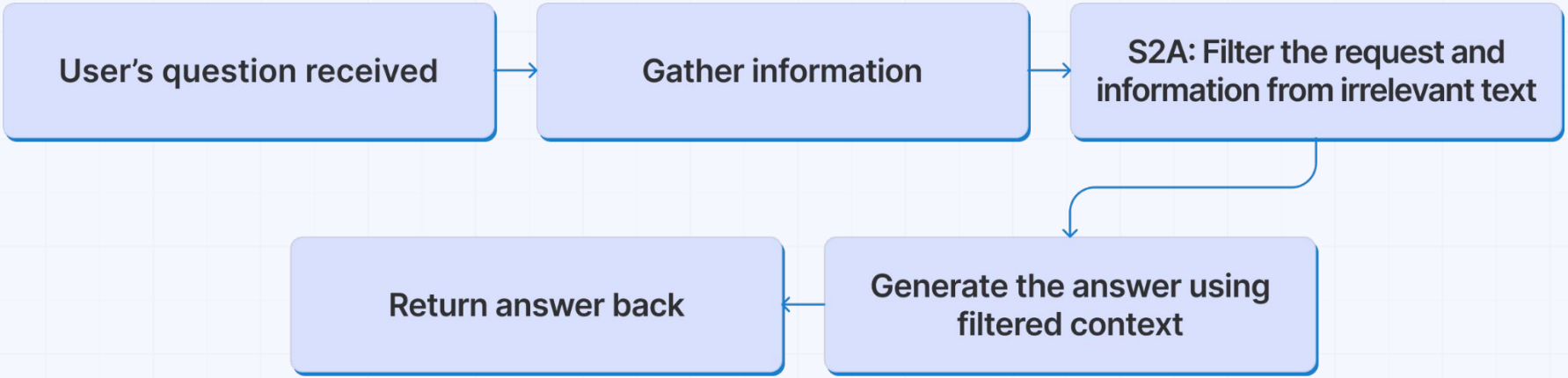
Want to dive deeper into this?

Feel free to check out these useful links for a more detailed scoop:

1. Academic Paper
2. GitHub Repository

Aleksei Kolesnikov
Staff Software Engineer

System 2 Attention (S2A) by Meta



Aleksei Kolesnikov
Staff Software Engineer

System 2 Attention (S2A) by Meta

Advantages

- → Improved performance
- → Better answer correlation
- → Huge improvements in case of biased or opinionated questions and requests

Additional details

Adding additional clean up steps will not be beneficial according to research

Disadvantages

- → The S2A approach often requires a large amount of training data
- → Can be computationally expensive to run
- → Still sometimes be affected by spurious correlations

*for more information you can visit: https://arxiv.org/pdf/2311.11829.pdf

Aleksei Kolesnikov
Staff Software Engineer

Reinforcement Learning from Human Feedback (RLHF)

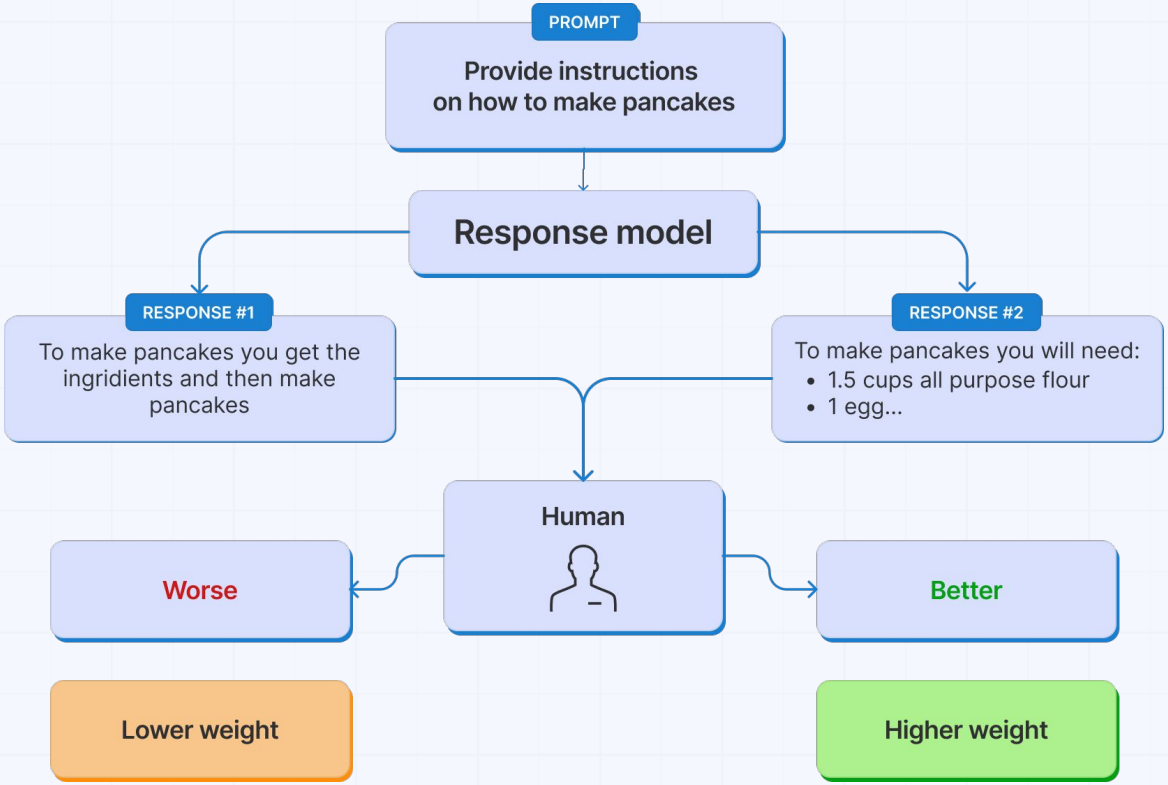
Idea

RLHF, used in AI assistants like ChatGPT, counts on human feedback for training via a Preference Model (PM).

This allows us to enhance answers and evaluate used sources based on user feedback.

For a deeper dive, you may visit the following repo:
https://github.com/Glareone/awesome-RLHF-GenAI

Reinforcement Learning from Human Feedback (RLHF)



Reinforcement Learning from Human Feedback (RLHF)

Advantages

- → Ethical alignment
- → **Context recognition**: Human evaluators can offer feedback specific to context

Limitations

- → Human bias
- → **Scalability issues**: The requirement for human input can limit RLHF's scalability

Aleksei Kolesnikov
Staff Software Engineer

RLAIF Reinforcement Learning from AI Feedback

Concept

RLAIF's utilizes another AI Feedback Model's feedback rather than direct human input.

Anthropic's RLAIF method has one AI model rectifying another based on a principles set, termed as "Constitutional AI". The application of these principles to AI judgments enables models to learn how to make better choices.

Where could be useful:

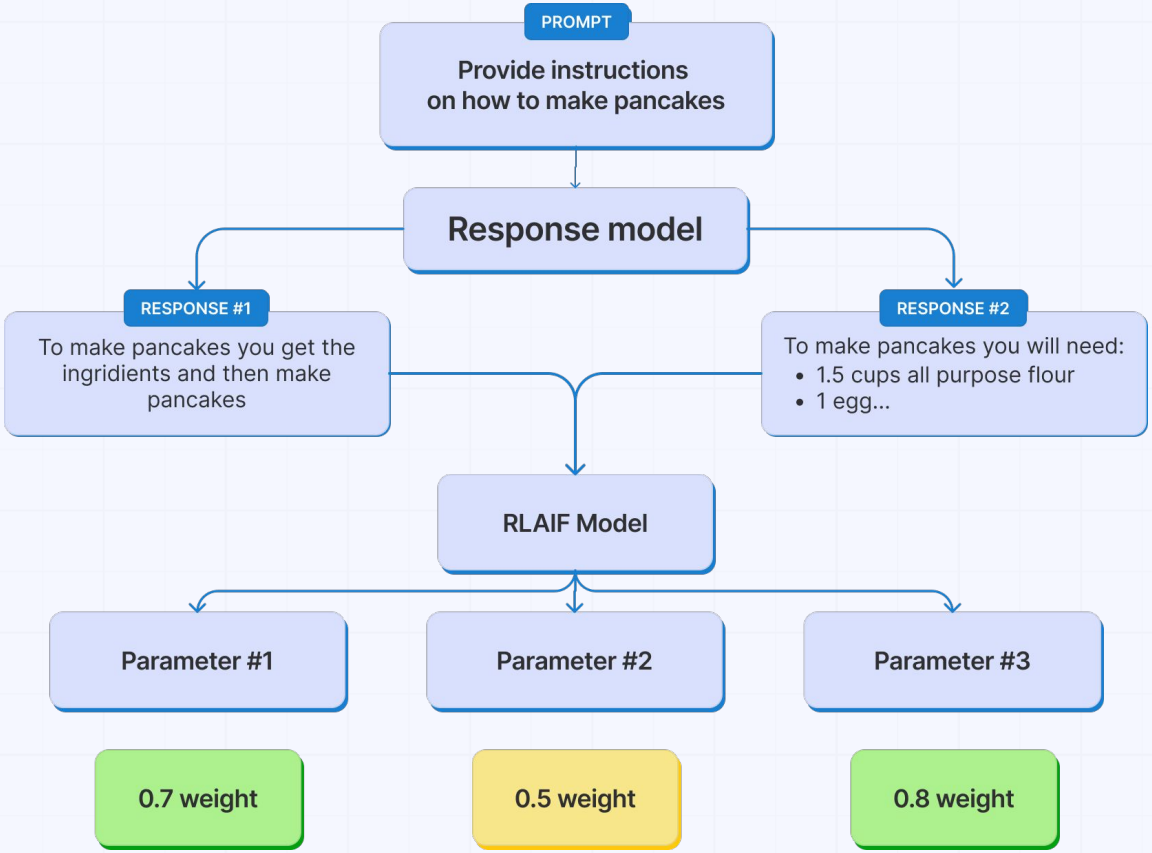
→ **Traffic Management:** AI systems can enhance traffic movement and lessen jams.

→ Ethnic and Compliance

→ **Environmental Monitoring:** AI-backed environmental tracking systems can process colossal data amounts to spot trends and give early hazard warnings.

Aleksei Kolesnikov
Staff Software Engineer

RLAIF Reinforcement Learning from AI Feedback



RLAIF Reinforcement Learning from AI Feedback

Advantages

→ **Scalability:** RLAIF surpasses RLHF in effective scalability

→ **Efficiency:** RLAIF fosters quicker learning and adaptation in AI systems

Limitations

→ **Prerequisite for robust principles:** RLAIF necessitates a broad set of rules

→ **Partiality:** AI-generated feedback may lack ethical examinations and nuances compared to human inputs.

For a deeper dive, you may visit the following repo: https://github.com/Glareone/awesome-RLHF-GenAI

Aleksei Kolesnikov
Staff Software Engineer

How to Handle Mixed Source-Type Information in GenAI?

Aleksei Kolesnikov
Staff Software Engineer

Imagine we have a variety of unstructured content like videos, text, images, and audio.

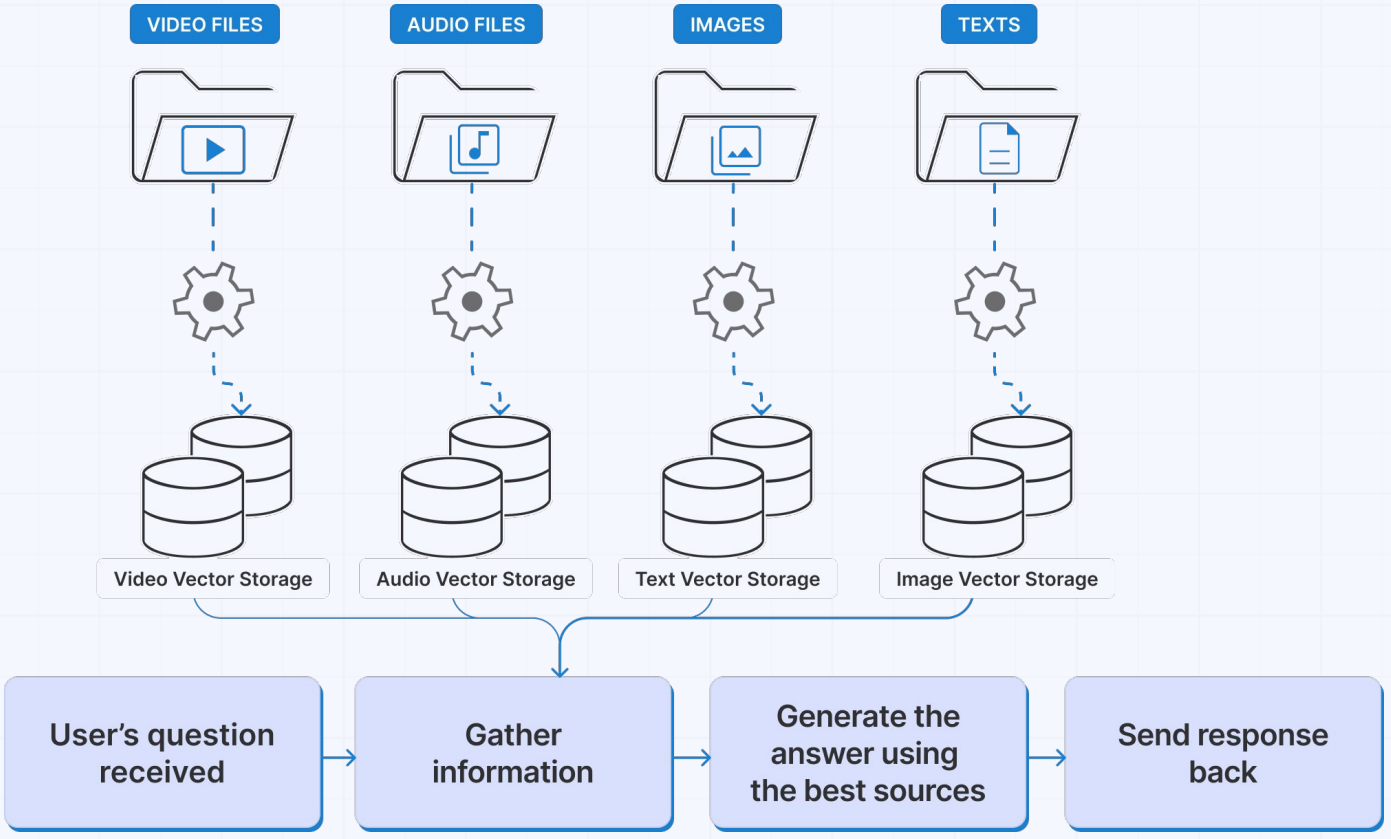
To answer queries, we need to draw information from all these sources.

So, how do we go about this?

There are two methods of achieving this:

1. The basic approach with **multiple vector storages**
2. The enhanced with one **multi-modal vector storage**

Basic Multi-Modal with Retrieval Augmented Generation



In a straightforward scenario, you can turn to embedding algorithms to create vectors from varied data and store them separately.

This raises a few questions:

1. Which algorithms can help extract information from images, audio, and videos?
2. Would using different embedding algorithms alter my responses?
3. Is it possible to consolidate vectors from different sources into one repository?

Aleksei Kolesnikov
Staff Software Engineer

Let's answer on them!

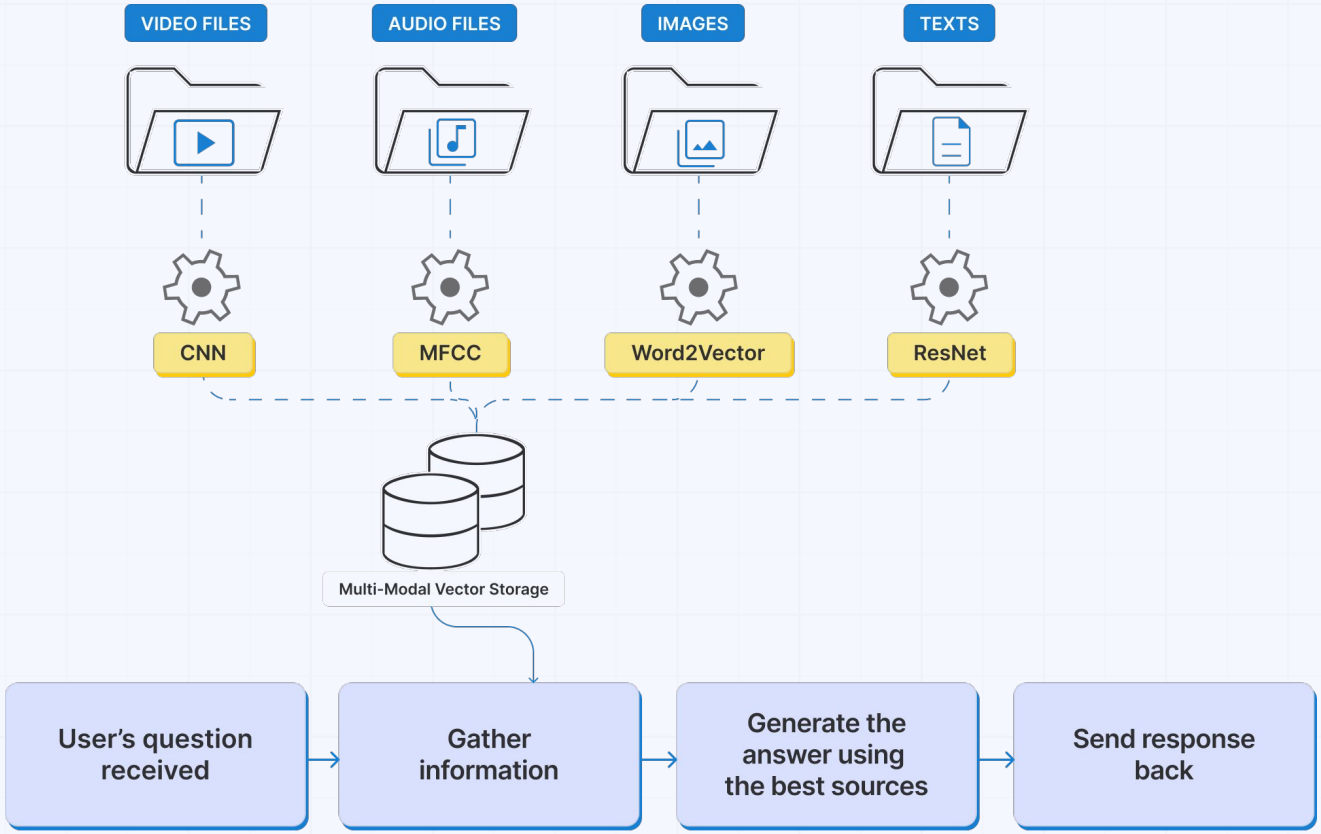
1. There is plenty of algorithms!
   - For text is quite obvious to go with **Word2Vec** or **BERT**,
   - For video it could be **Convolutional Neural Network** (**CNN**) or Supervised Contrastive Learning (**CNN-SCL**).
   - For audio files the most commonly used is **MFCC** (**Mel Frequency Cepstral Coefficients**).
   - For images **ResNet** could be used.

2. Of course it affects genAI responses! Using different algorithms can result in different embeddings, leading to varying responses generated by the multi-modal RAG paradigm!

3. Definitely Possible using Multi-modal index! By creating a combined representation for each piece of multi-modal data, search and retrieval processes can be greatly enhanced. **But it's a challenging process.**

Recommended articles:
https://arxiv.org/pdf/2306.08789.pdf
https://arxiv.org/pdf/2104.08108.pdf

Basic Multi-Modal with Retrieval Augmented Generation



How to scan newly added files? Push or Pull (Scan) strategies!

→ With the **Push approach**, event-triggered approach could be leveraged, such that each time a new file is added, it's automatically processed into a vector.

→ Alternatively, using the **Pull approach**, scanning process could be setup and it will process files in the storage at periodic intervals, like every hour, for instance. Used in Semantic Search.

Aleksei Kolesnikov
Staff Software Engineer

The final question might be the following

Can the Multi-modal RAG approach function synergistically with other RAG or FLARE paradigms?

The response is YES!

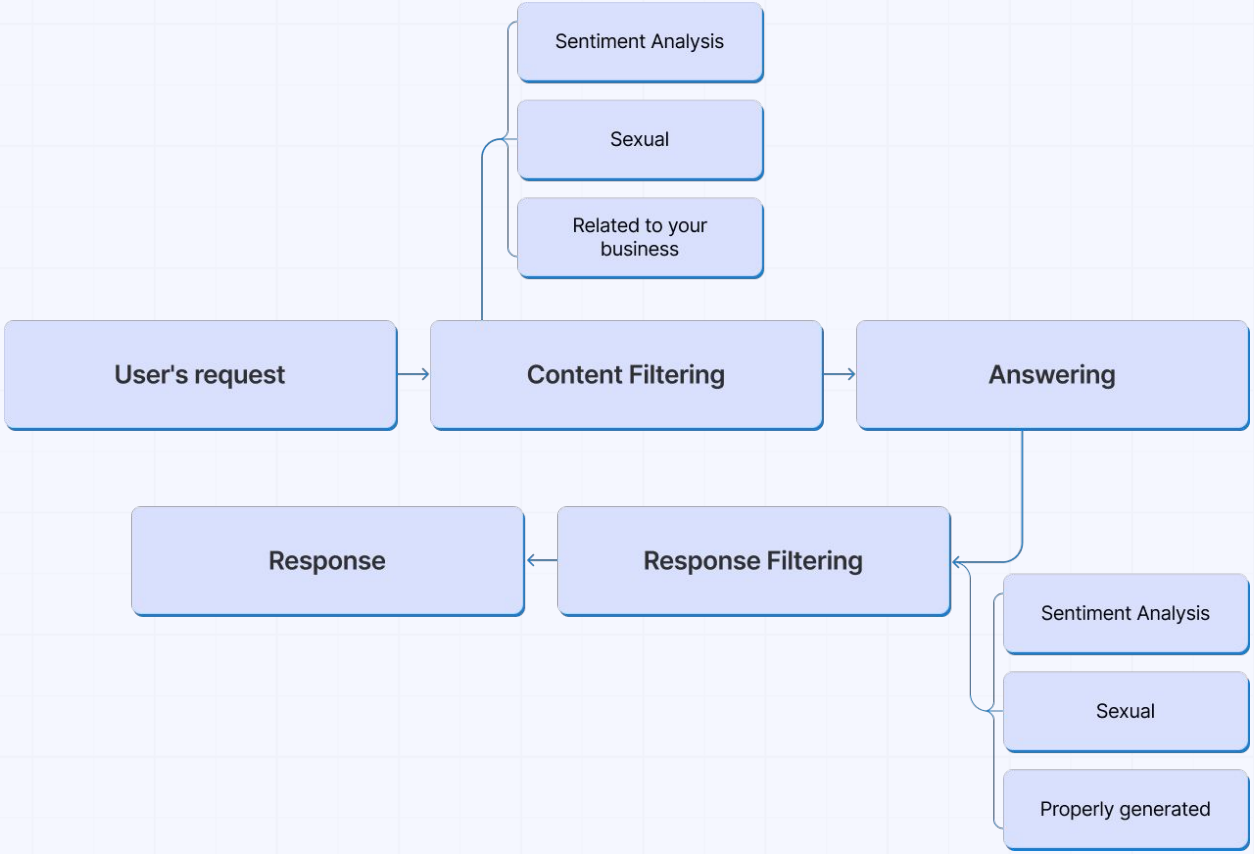
Numerous articles and statistics attest to enhanced accuracy and relevance in responses. However, be aware that this could come with a considerable performance cost and seamless integration issues.

Good article was published on llamaindex:
https://docs.llamaindex.ai/en/stable/examples/evaluation/multi_modal/multi_modal_rag_evaluation.html

Content Filtering

Aleksei Kolesnikov
Staff Software Engineer

Content Filtering



Thank you!
Questions?

Aleksei Kolesnikov
Staff Software Engineer