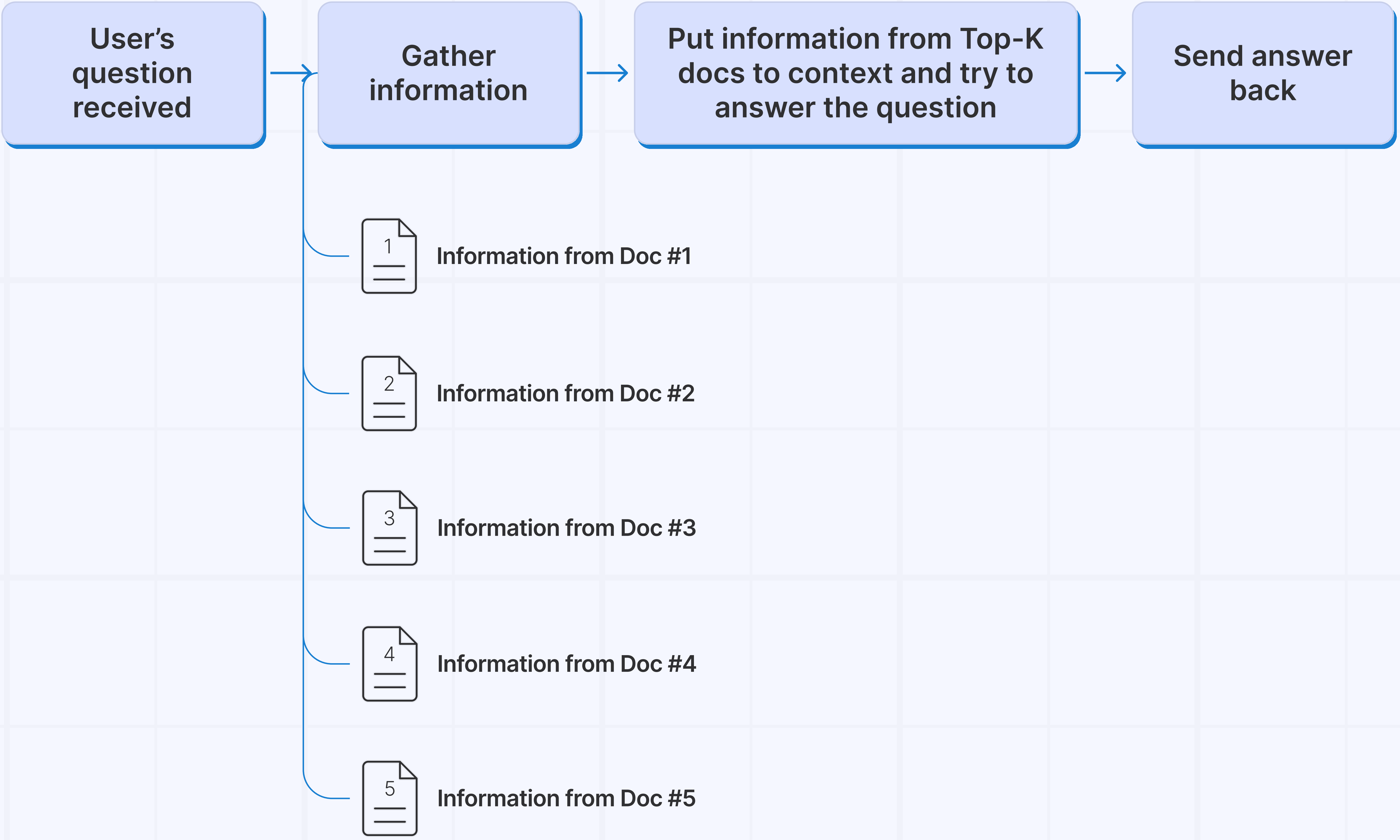
Enhancing Answer Relevance: Straightforward Paradigms for Improvement

Aleksei Kolesnikov

Staff Software Engineer

RAG paradigm using Top-K wording in simple words



User's question received

Gather information

Put information from Top-K docs to context and try to answer the question

Send answer back

Information from Doc #1

Information from Doc #2

Information from Doc #3

Information from Doc #4

Information from Doc #5

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→ What if information is not relevant or outdated?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!

Re-ranking in RAG paradigm

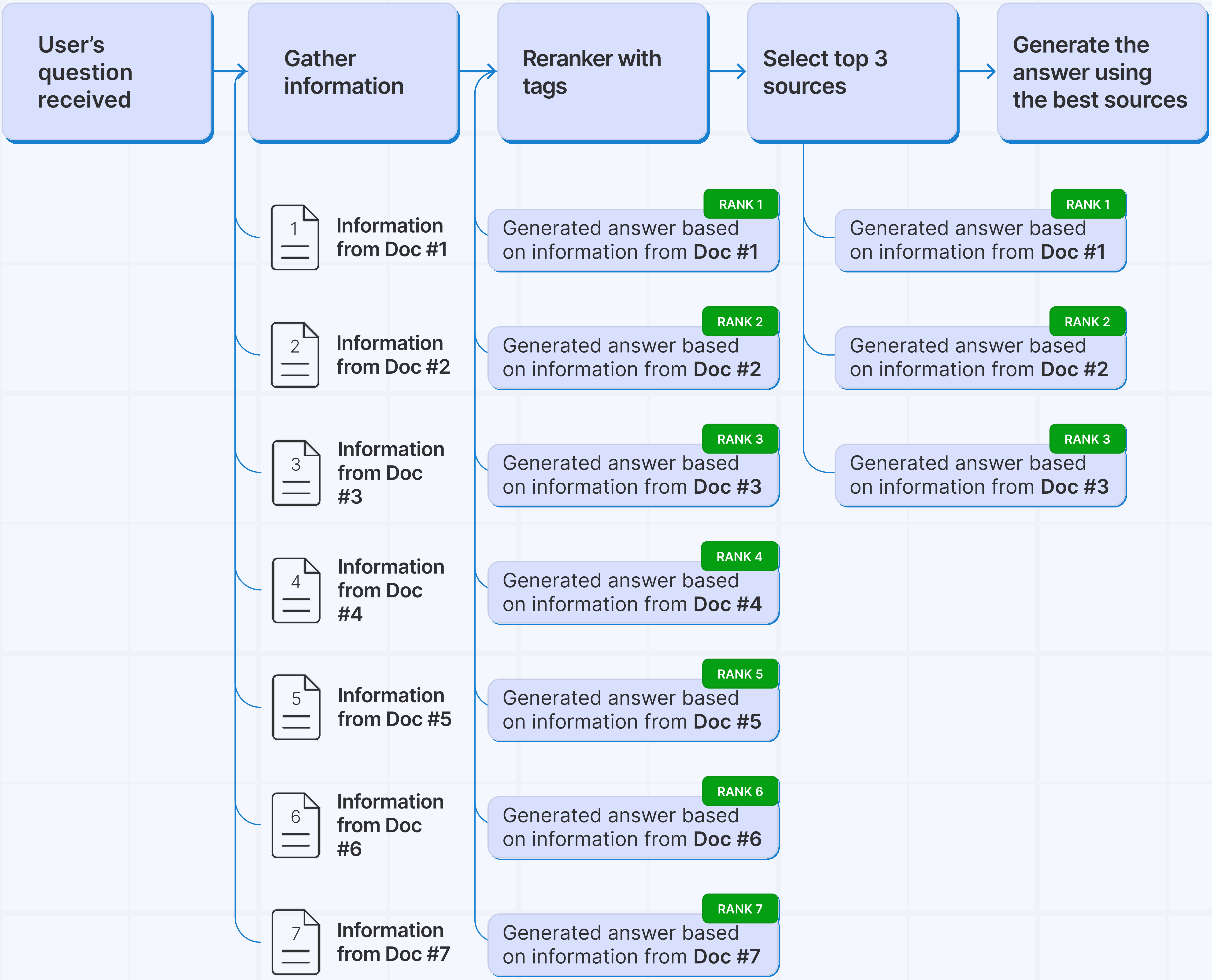
Idea:

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.

Re-ranking in RAG



User's question received → Gather information → Reranker with tags → Select top 3 sources → Generate the answer using the best sources

Gather information:

- Information from Doc #1
- Information from Doc #2
- Information from Doc #3
- Information from Doc #4
- Information from Doc #5
- Information from Doc #6
- Information from Doc #7

Reranker with tags:

- RANK 1 Generated answer based on information from **Doc #1**
- RANK 2 Generated answer based on information from **Doc #2**
- RANK 3 Generated answer based on information from **Doc #3**
- RANK 4 Generated answer based on information from **Doc #4**
- RANK 5 Generated answer based on information from **Doc #5**
- RANK 6 Generated answer based on information from **Doc #6**
- RANK 7 Generated answer based on information from **Doc #7**

Select top 3 sources:

- RANK 1 Generated answer based on information from **Doc #1**
- RANK 2 Generated answer based on information from **Doc #2**
- RANK 3 Generated answer based on information from **Doc #3**

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]

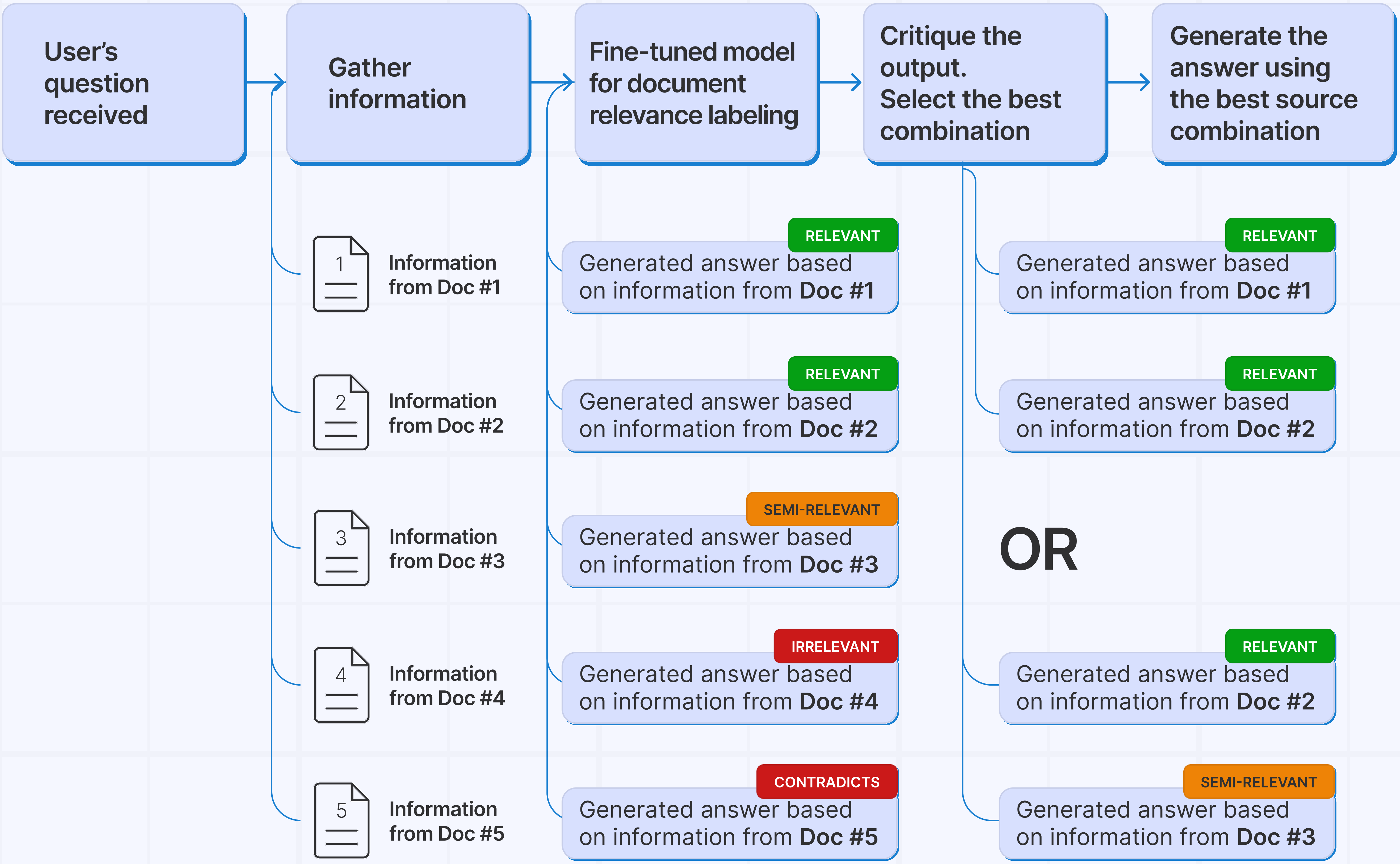
Self-RAG operates through a sequential process

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/

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

Information from Doc #1

Information from Doc #2

Information from Doc #3

Information from Doc #4

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.

FLARE
Forward-Looking Active Retrieval Augmented Generation

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.

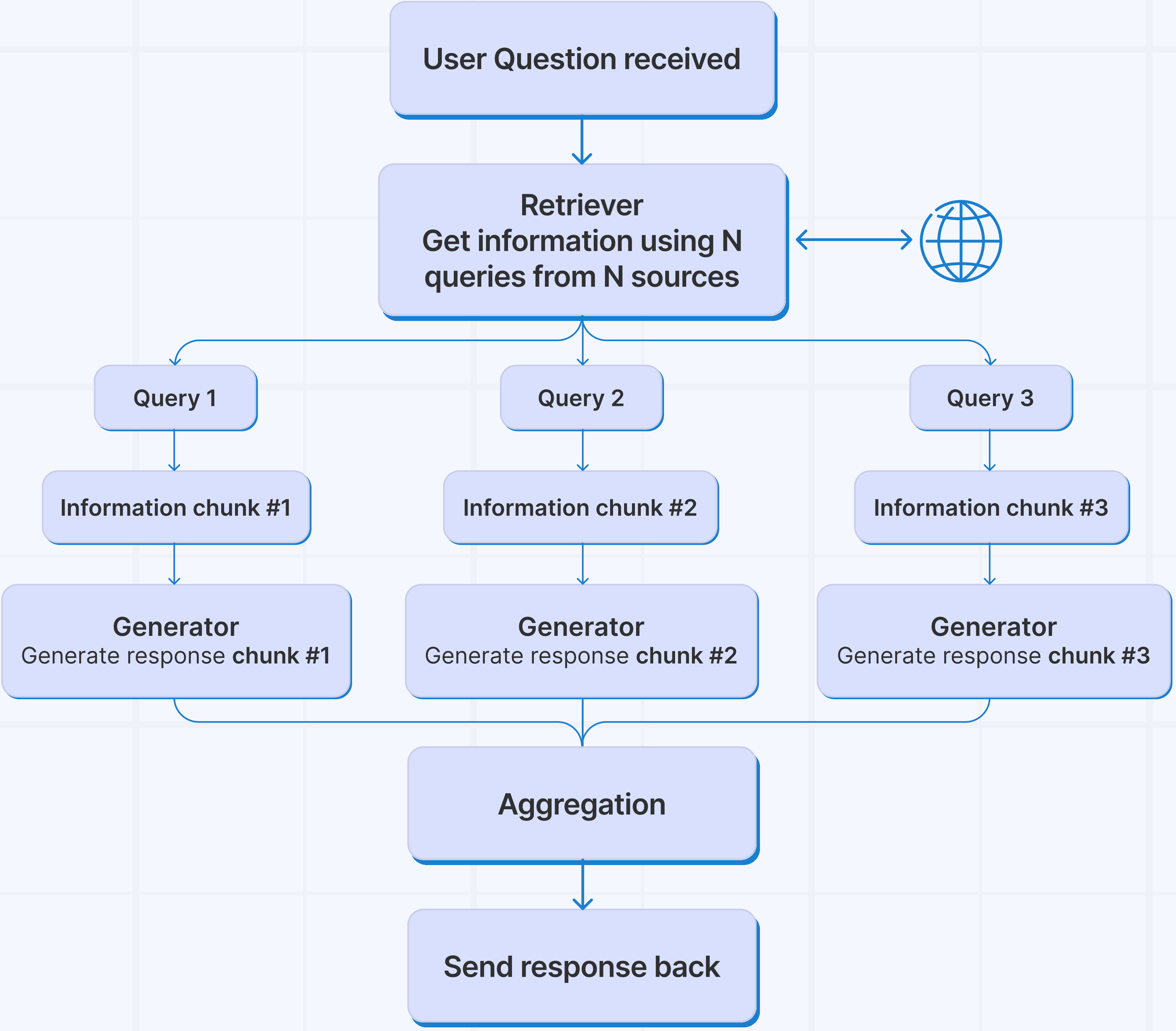
Forward-Looking Active Retrieval Augmented Generation (FLARE) Advantages

→ Works good if you need a summary

→ Can aggregate information from several sources

→ Can answer on questions to predict next steps.

Forward-Looking Active Retrieval Augmented Generation (FLARE)



User Question received

Retriever
Get information using N queries from N sources

Query 1

Query 2

Query 3

Information chunk #1

Information chunk #2

Information chunk #3

Generator
Generate response **chunk #1**

Generator
Generate response **chunk #2**

Generator
Generate response **chunk #3**

Aggregation

Send response back

Aleksei Kolesnikov
Staff Software Engineer