ChatGT3

What's up?

Improving Retrieval Mechanism in Retrieval-Augmented Generation (RAG) Architecture

Applied Data Science Project 2024

Enhanced Retrieval
Smarter Responses

Project name*

ADSP – P9 – RAG MARCO

Continue

OR CONTACT

Homayoun Afshari

Arash Daneshvar

Hossein Khodadadi

ChatGT3

ChatGT3

Explore GT3s

Today

- Project Objectives
- Problem Statement
- Quick Solution
- Methodology: Introduction
- Methodology: Metrics
- Methodology: Baselines
- Proposed Method
- Comparisons
- Conclusion

Enhanced Retrieval
Smarter Responses

What can I help with?

Message ChatGT3

Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

ChatGT3

Explore GT3s

Today

Project Objectives

Problem Statement

Quick Solution

Methodology: Introduction

Methodology: Metrics

Methodology: Baselines

Proposed Method

Comparisons

Conclusion

Enhanced Retrieval

Smarter Responses

ChatGT3

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What are the values of the system described in the attached PDFs?

Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Project Objectives

The objectives of this study are as follows:

- **OBJECTIVE 1**: Enhance the **Retrieval Mechanism** by leveraging SOTA techniques proposed by the literature.
- **OBJECTIVE 2**: Enhance the **Evaluation Metrics** of the retrieved documents to provide reliable context for the LLM.

Furthermore, the alignment with the united nations Sustainable Development Goals (SDGs), the project could relate to the following items:

- **SDG 4 (QUALITY EDUCATION)**: The project improves information access, supporting quality education through enhanced knowledge retrieval.
- **SDG 9 (INDUSTRY, INNOVATION, AND INFRASTRUCTURE)**: By advancing retrieval technology, the project promotes innovation and strengthens information infrastructure.

ChatGT3

ChatGT3

Explore GT3s

Today

Project Objectives

Problem Statement

Quick Solution

Methodology: Introduction

Methodology: Metrics

Methodology: Baselines

Proposed Method

Comparisons

Conclusion

Enhanced Retrieval

Smarter Responses

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What problem does this project aim to solve based on the attached PDF?

Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

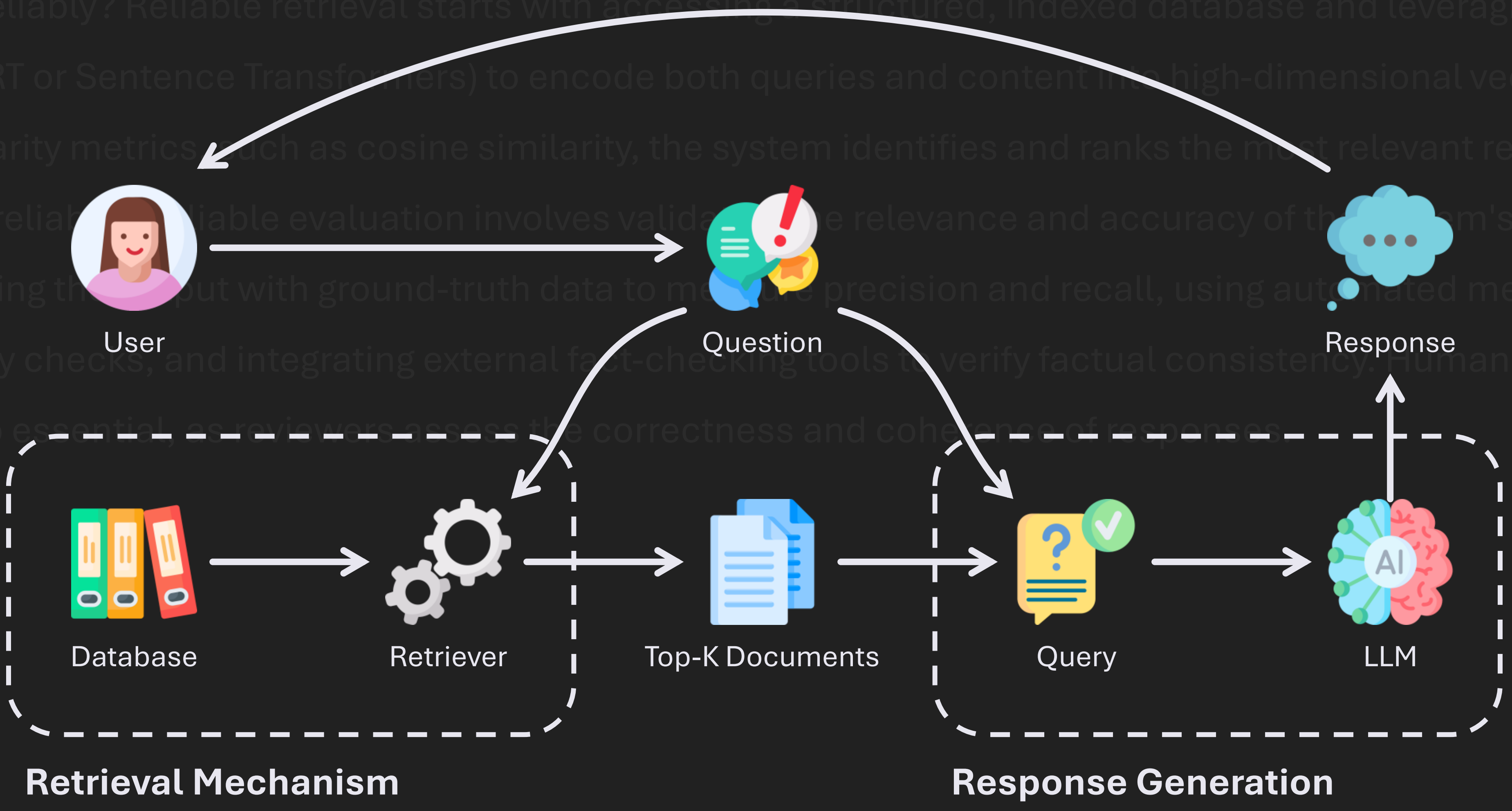
Problem Statement

The RAG (Retrieve-and-Generate) system can be explained by addressing two key questions:

"How to retrieve reliably?" and **"How to evaluate reliably?"**

How to retrieve reliably? Reliable retrieval starts with accessing a structured, indexed database and leveraging embedding models (e.g., BERT or Sentence Transfo[illegible]ers) to encode both queries and content [illegible] high-dimensional vectors. By calculating similarity metrics [illegible]uch as cosine similarity, the system identifies and ranks the m[illegible]t relevant results.

How to evaluate relia[illegible]liable evaluation involves valida[illegible]e relevance and accuracy of th[illegible]m's responses. This includes comparing th[illegible]put with ground-truth data to m[illegible] precision and recall, using aut[illegible]ated metrics like BLEU or ROUGE for quality checks, and integrating external fa[illegible]t-checking tools t[illegible] verify factual consistency. Human-in-the-loop evaluation is also essential, as reviewers assess th[illegible] correctness and coherence of responses.



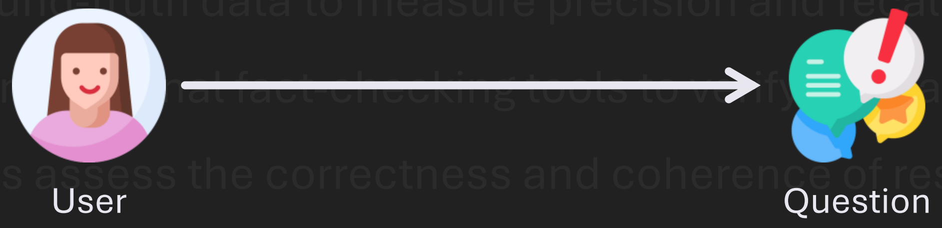
Problem Statement

The RAG (Retrieve-and-Generate) system can be explained by addressing two key questions:

"How to retrieve reliably?" and **"How to evaluate reliably?"**

How to retrieve reliably? Reliable retrieval starts with accessing a structured, indexed database and leveraging embedding models (e.g., BERT or Sentence Transformers) to encode both queries and content into high-dimensional vectors. By calculating similarity metrics, such as cosine similarity, the system identifies and ranks the most relevant results.

How to evaluate reliably? Reliable evaluation involves validating the relevance and accuracy of the system's responses. This includes comparing the output with ground-truth data to measure precision and recall, using automated metrics like BLEU or ROUGE for quality checks, and integrati… al fact-checking tools to v… al consistency. Human-in-the-loop evaluation is also essential, as reviewers assess the correctness and coherence of responses.



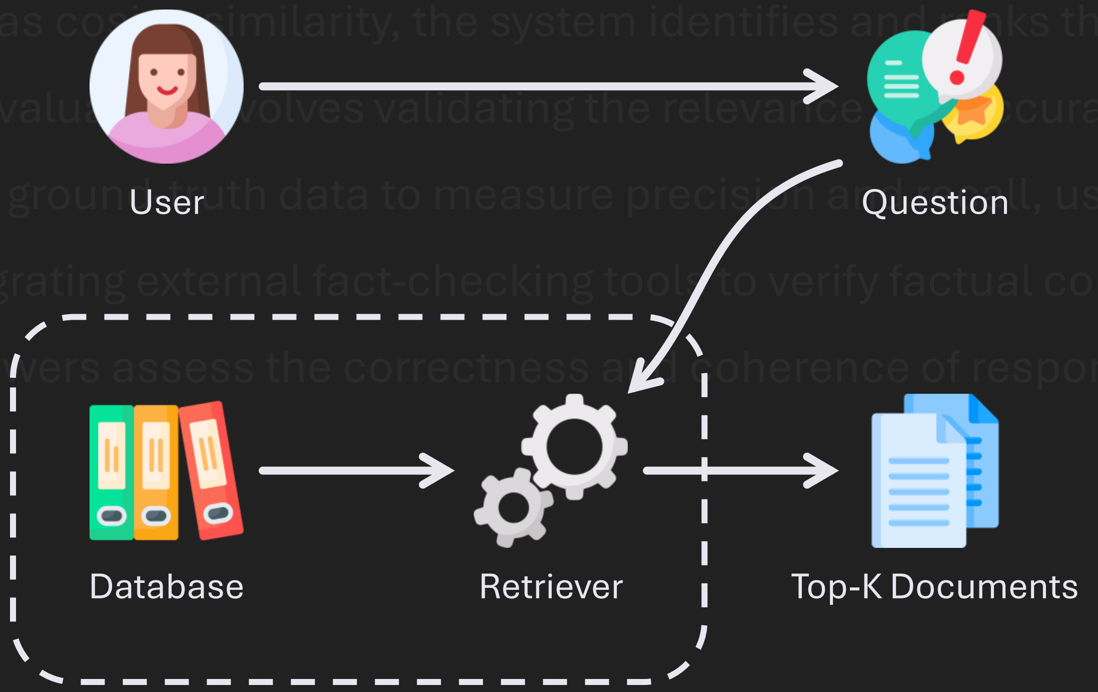
Problem Statement

The RAG (Retrieve-and-Generate) system can be explained by addressing two key questions:

"How to retrieve reliably?" and **"How to evaluate reliably?"**

How to retrieve reliably? Reliable retrieval starts with accessing a structured, indexed database and leveraging embedding models (e.g., BERT or Sentence Transformers) to encode both queries and content into high-dimensional vectors. By calculating similarity metrics, such as cosine similarity, the system identifies and ranks the most relevant results.

How to evaluate reliably? Reliable evaluation involves validating the relevance and accuracy of the system's responses. This includes comparing the output with ground truth data to measure precision and recall, using automated metrics like BLEU or ROUGE for quality checks, and integrating external fact-checking tools to verify factual consistency. Human-in-the-loop evaluation is also essential, as reviewers assess the correctness and coherence of responses.

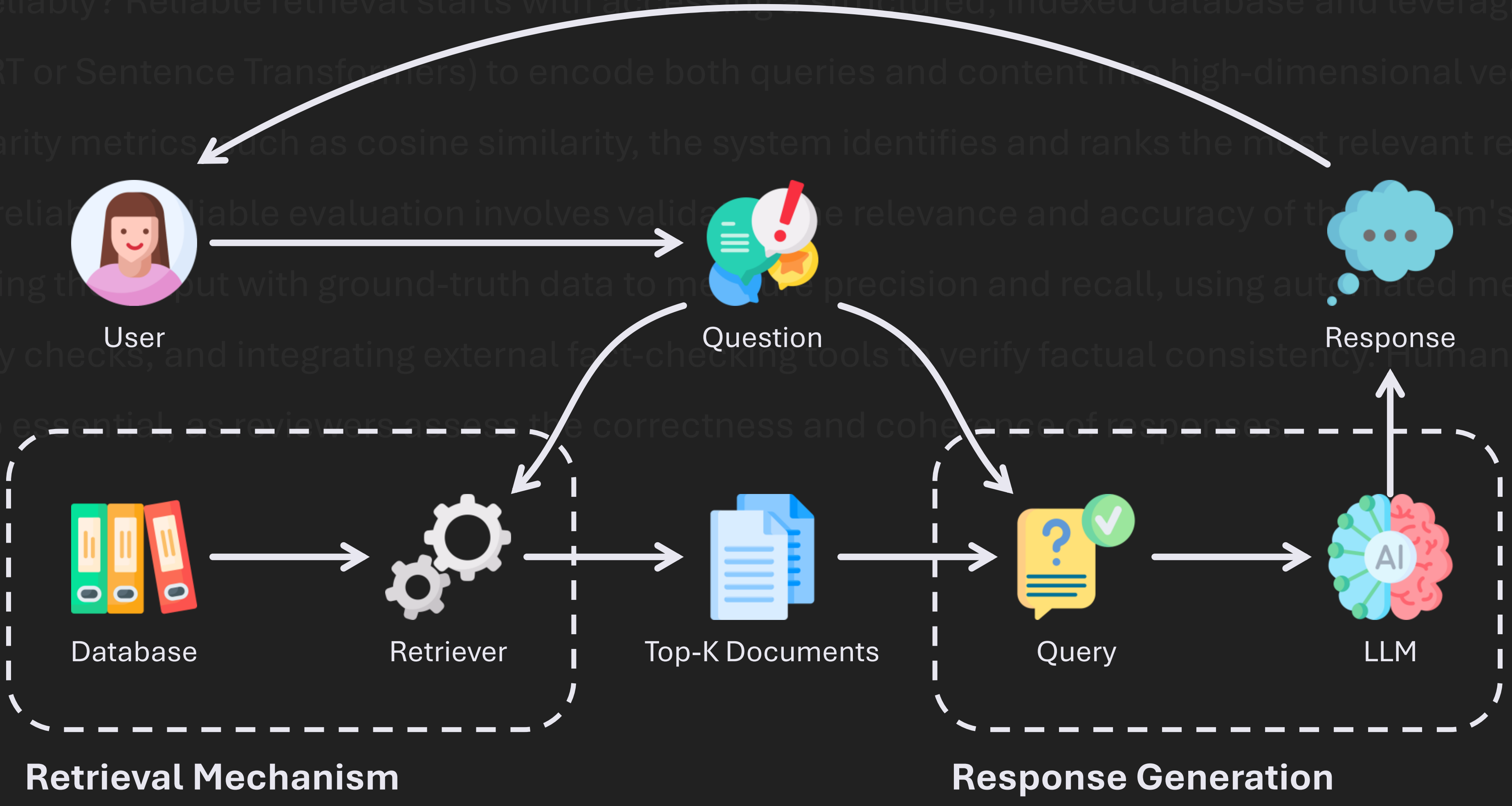


Retrieval Mechanism

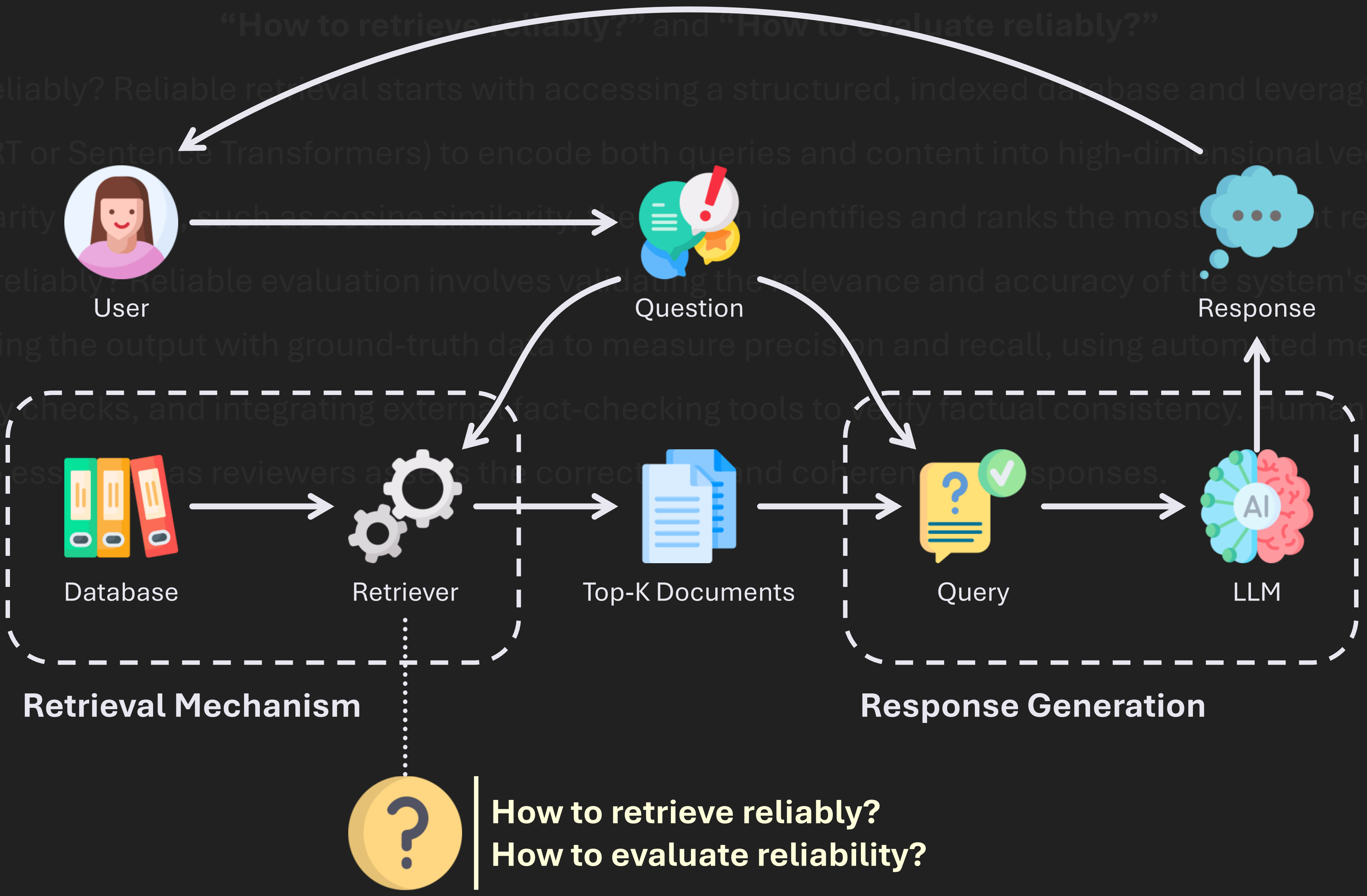
Problem Statement

The RAG (Retrieve-and-Generate) system can be explained by addressing two key questions:

“How to retrieve reliably?” and **“How to evaluate reliably?”**



Problem Statement



ChatGT3

Explore GT3s

Today

Project Objectives

Problem Statement

Quick Solution

Methodology: Introduction

Methodology: Metrics

Methodology: Baselines

Proposed Method

Comparisons

Conclusion

Enhanced Retrieval
Smarter Responses

ChatGT3

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What kind of achievement is described in the attached PDFs?

Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

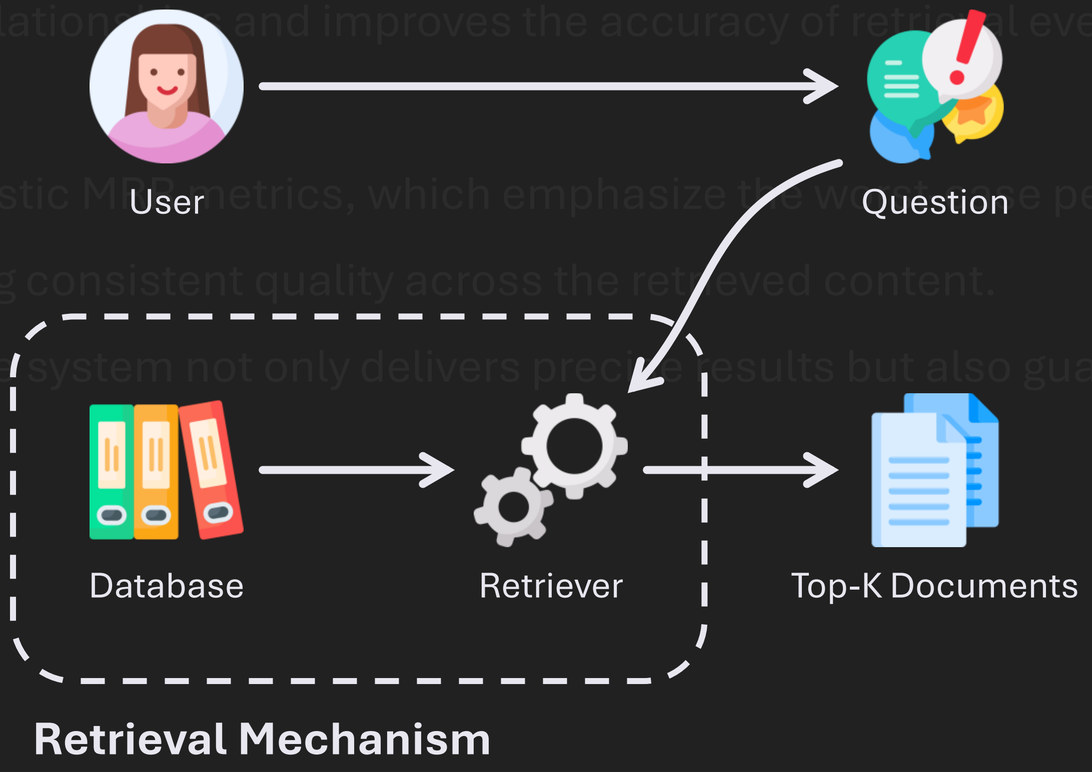
Quick Solution

Our solution to address these two questions—how to retrieve reliably and how to evaluate reliably—is by leveraging semantic search using a mapper and pessimistic Mean Reciprocal Rank (MRR) metrics.

Semantic search ensures reliable retrieval by encoding both queries and documents into high-dimensional vectors using a mapper, which aligns the query intent with the most relevant content based on semantic meaning rather than just keywords. This method captures contextual relationships and improves the accuracy of retrieval even for complex or ambiguous queries.

To evaluate reliably, we use pessimistic MRR metrics, which emphasize the worst-case performance by penalizing low-ranked but relevant results, ensuring consistent quality across the retrieved content.

By combining these approaches, the system not only delivers precise results but also guarantees robustness under varying query scenarios.



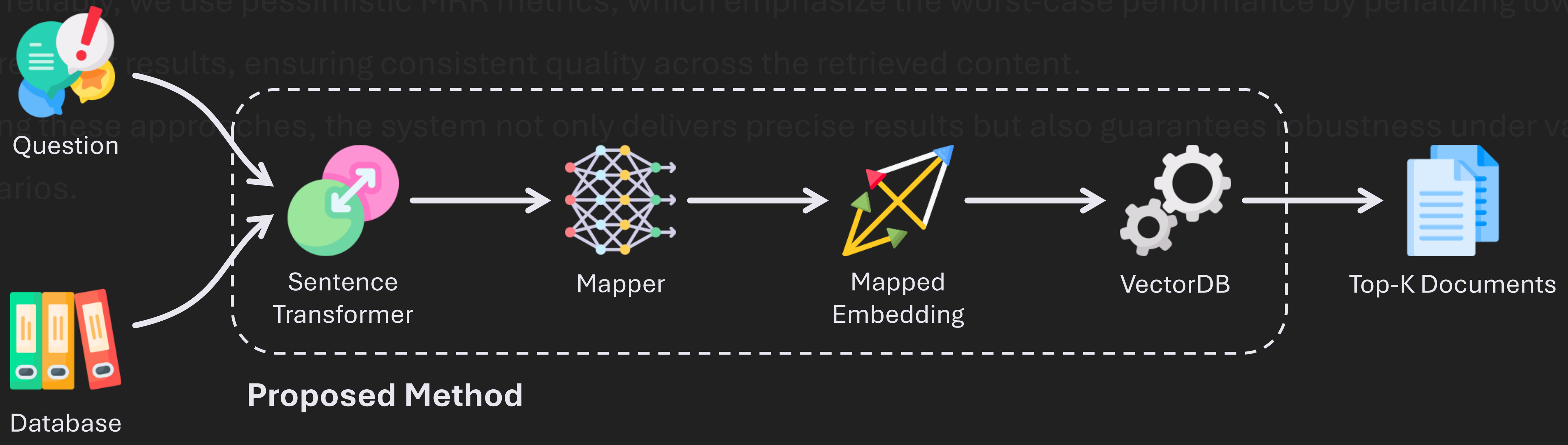
Quick Solution

Our solution to address these two questions—how to retrieve reliably and how to evaluate reliably—is by leveraging semantic search using a mapper and pessimistic Mean Reciprocal Rank (MRR) metrics.

Semantic search ensures reliable retrieval by encoding both queries and documents into high-dimensional vectors using a mapper, which aligns the query intent with the most relevant content based on semantic meaning rather than just keywords. This method captures contextual relationships and improves the accuracy of retrieval even for complex or ambiguous queries.

To evaluate reliably, we use pessimistic MRR metrics, which emphasize the worst-case performance by penalizing low-ranked but r… results, ensuring consistent quality across the retrieved content.

By combining these approaches, the system not only delivers precise results but also guarantees robustness under varying query scenarios.



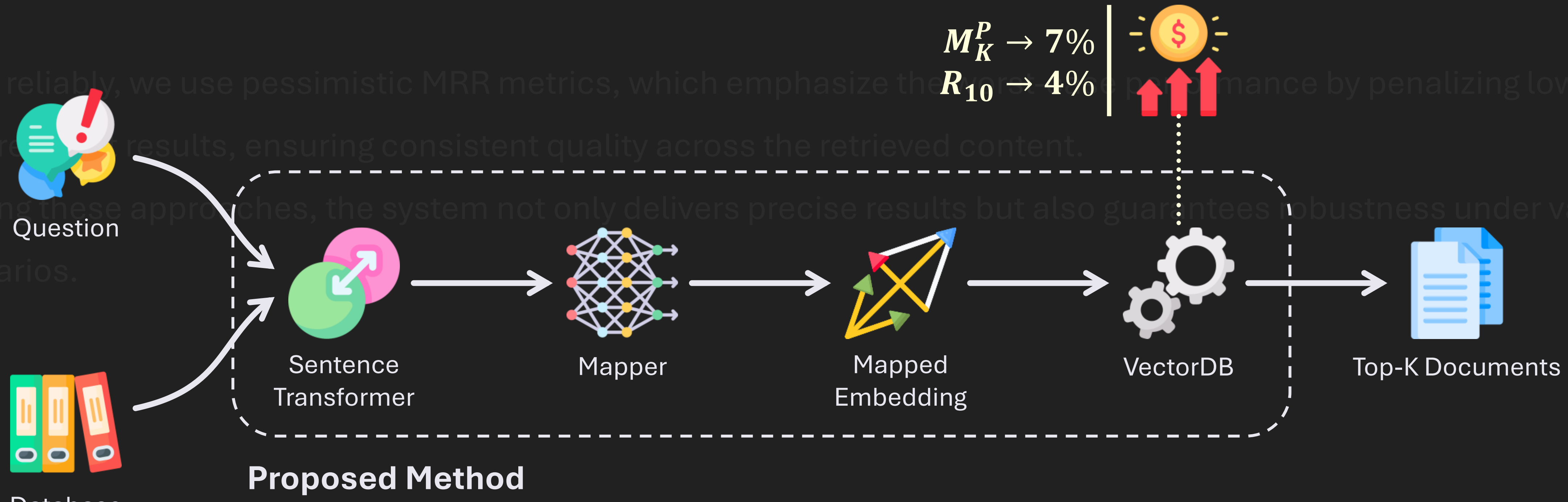
Quick Solution

Our solution to address these two questions—how to retrieve reliably and how to evaluate reliably—is by leveraging semantic search using a mapper and pessimistic Mean Reciprocal Rank (MRR) metrics.

Semantic search ensures reliable retrieval by encoding both queries and documents into high-dimensional vectors using a mapper, which aligns the query intent with the most relevant content based on semantic meaning rather than just keywords. This method captures contextual relationships and improves the accuracy of retrieval even for complex or ambiguous queries.

To evaluate reliably, we use pessimistic MRR metrics, which emphasize the [illegible] performance by penalizing low-ranked but r[illegible] results, ensuring consistent quality across the retrieved content.

By combining these approaches, the system not only delivers precise results but also guarantees robustness under varying query scenarios.



ChatGT3

ChatGT3

Explore GT3s

Today

Project Objectives

Problem Statement

Quick Solution

Methodology: Introduction

Methodology: Metrics

Methodology: Baselines

Proposed Method

Comparisons

Conclusion

Enhanced Retrieval
Smarter Responses

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

How did we arrive at this result and what challenges did we face along the way?

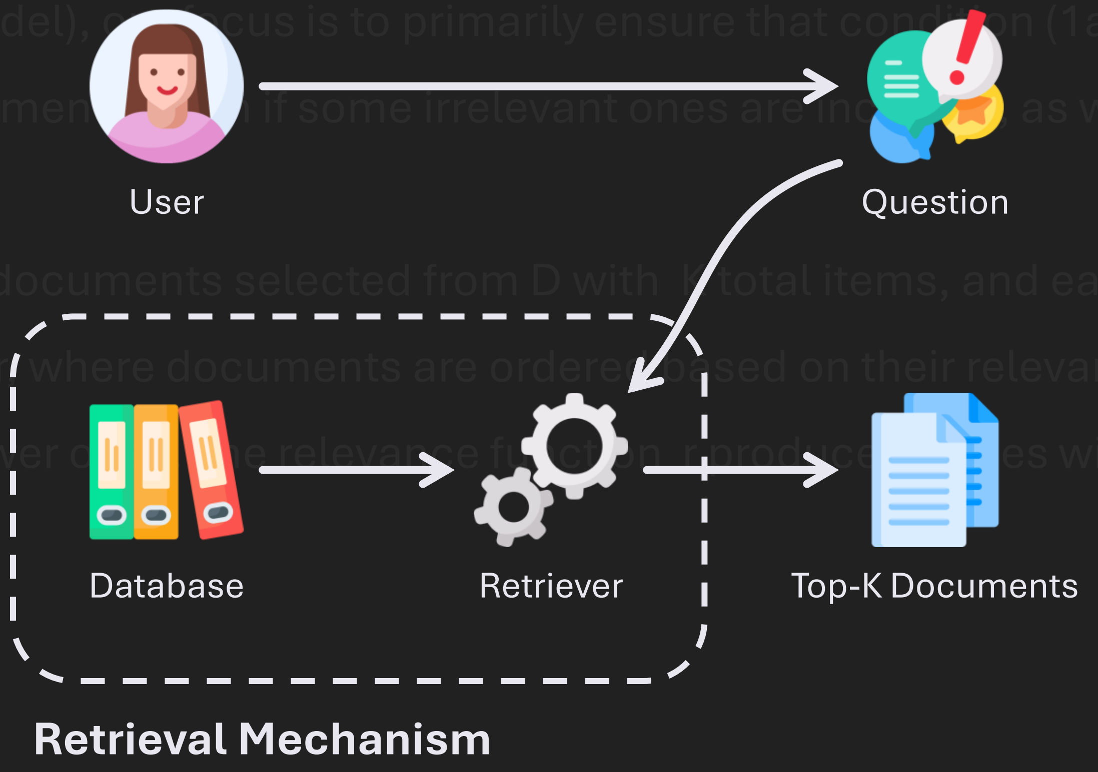
Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Methodology: Introduction

The first step in our methodology for this project is to define the primary task, which involves designing a retrieval mechanism to obtain an optimal set of documents, RK (q), that satisfies two key conditions: all relevant documents are retrieved while no irrelevant ones are included. These conditions, expressed as $G(q) \subseteq RK(q)$ and $RK(q) \subseteq G(q)$, ensure the precision of the retrieval process. However, given that the next step in a RAG (Retrieval-Augmented Generation) system is a response-generating LLM (Large Language Model), o[illegible]cus is to primarily ensure that cond[illegible]n (1a) is satisfied. This means we prioritize retrieving all relevant documen[illegible] if some irrelevant ones are inc[illegible] as we can rely on the LLM's capability to handle the redundancy.

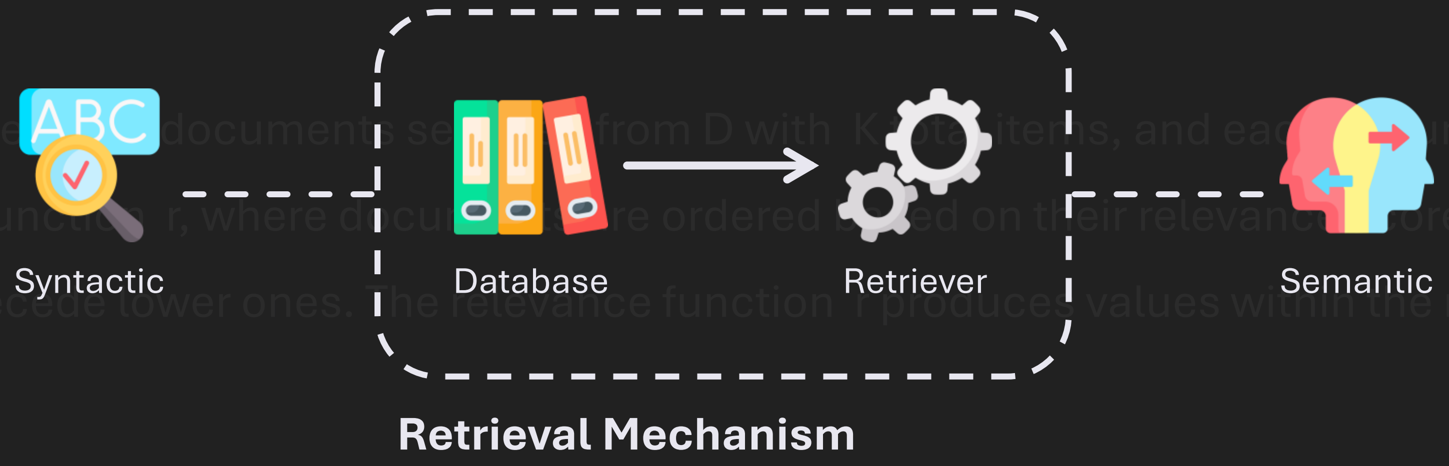
Thus, we define RK (q) as the set of documents selected from D with K total items, and each document's relevance is measured using a ranking function r, where documents are ordere[illegible]ased on their relevance score $r(q,d_i)$, ensuring that higher relevance scores precede lo[illegible] relevance f[illegible]ction [illegible] produce[illegible]es within the range [0,1].



Methodology: Introduction

The first step in our methodology for this project is to define the primary task, which involves designing a retrieval mechanism to obtain an optimal set of documents, RK (q), that satisfies two key conditions: all relevant documents are retrieved while no irrelevant ones are included. These conditions, expressed as $G(q) \subseteq RK(q)$ and $RK(q) \subseteq G(q)$, ensure the precision of the retrieval process. However, given that the next step in a RAG (Retrieval-Augmented Generation) system is a response-generating LLM (Large Language Model), our focus is to primarily ensure that condition (1a) is satisfied. This means we prioritize retrieving all relevant documents, even if some irrelevant ones are included, as we can rely on the LLM's capability to handle the redundancy.

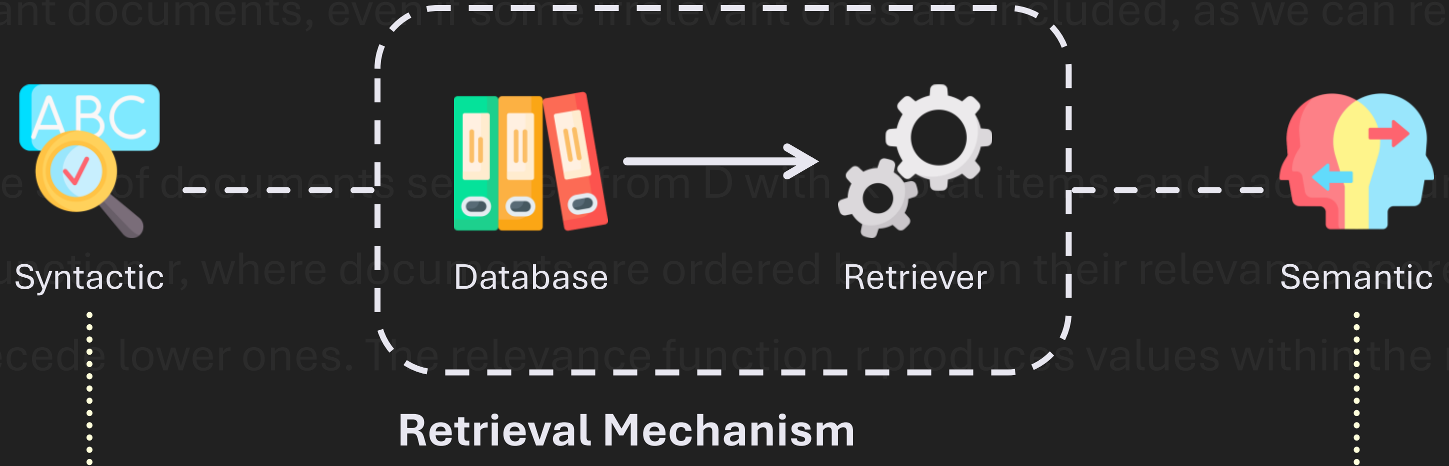
Thus, we define RK (q) as the [illegible] documents se[illegible] from D with K [illegible] items, and ea[illegible]ment's relevance is measured using a ranking fun[illegible] r, where docu[illegible] re ordered [illegible] ed on their relevanc[illegible] re r(q,di), ensuring that higher relevance scores prec[illegible] lower ones. The relevance function r produces values within the range [0,1].



Methodology: Introduction

The first step in our methodology for this project is to define the primary task, which involves designing a retrieval mechanism to obtain an optimal set of documents, RK (q), that satisfies two key conditions: all relevant documents are retrieved while no irrelevant ones are included. These conditions, expressed as G(q)⊆RK (q) and RK (q)⊆G(q), ensure the precision of the retrieval process. However, given that the next step in a RAG (Retrieval-Augmented Generation) system is a response-generating LLM (Large Language Model), our focus is to primarily ensure that condition (1a) is satisfied. This means we prioritize retrieving all relevant documents, even if some irrelevant ones are included, as we can rely on the LLM's capability to handle the redundancy.

Thus, we define RK (q) as the [illegible] of documents se[illegible] from D with [illegible] items, and ea[illegible]ment's relevance is measured using a ranking fu[illegible], where do[illegible] are ordered [illegible] their relevan[illegible] r(q,di), ensuring that higher relevance scores precede lower ones. The relevance function r produces values within the range [0,1].



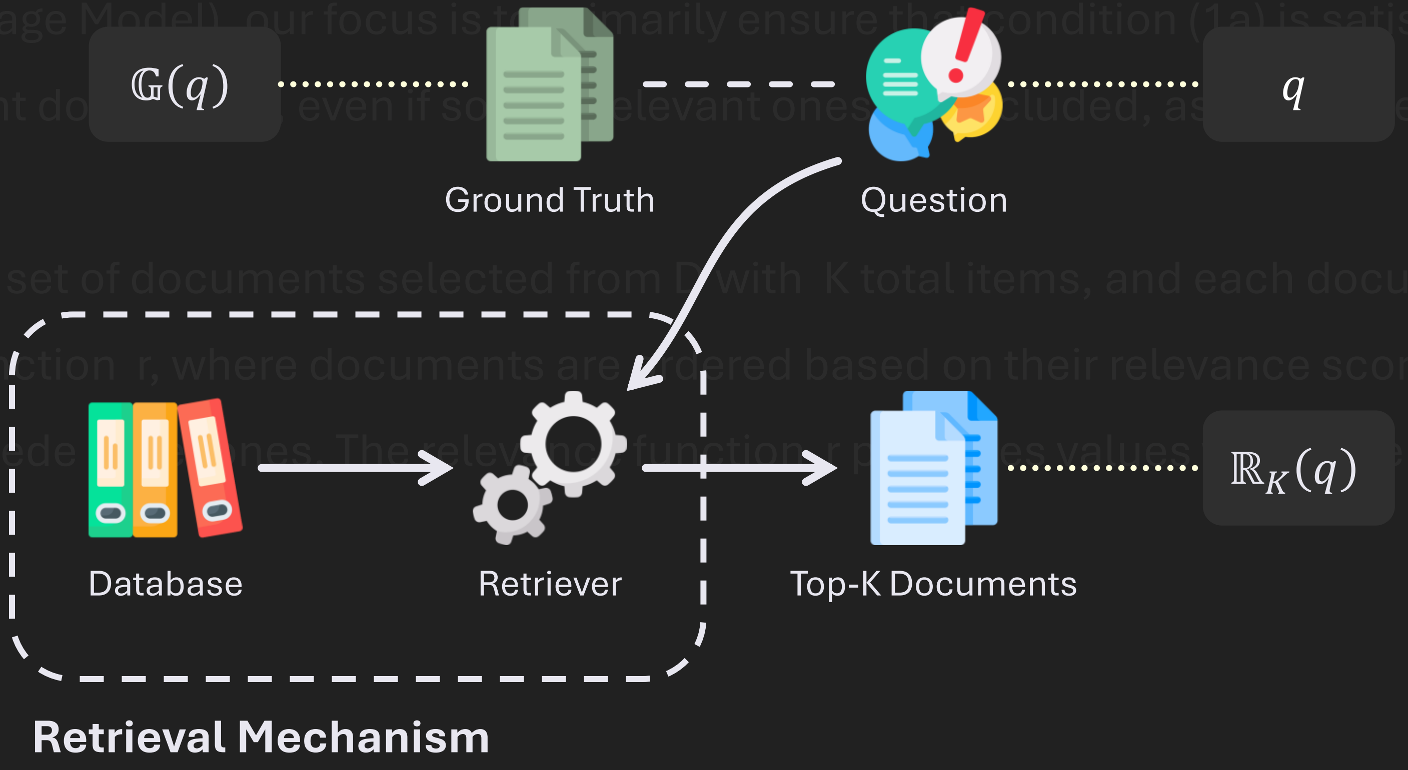
Carnegie Mellon
Hotpot-QA

Microsoft
MS-Marco

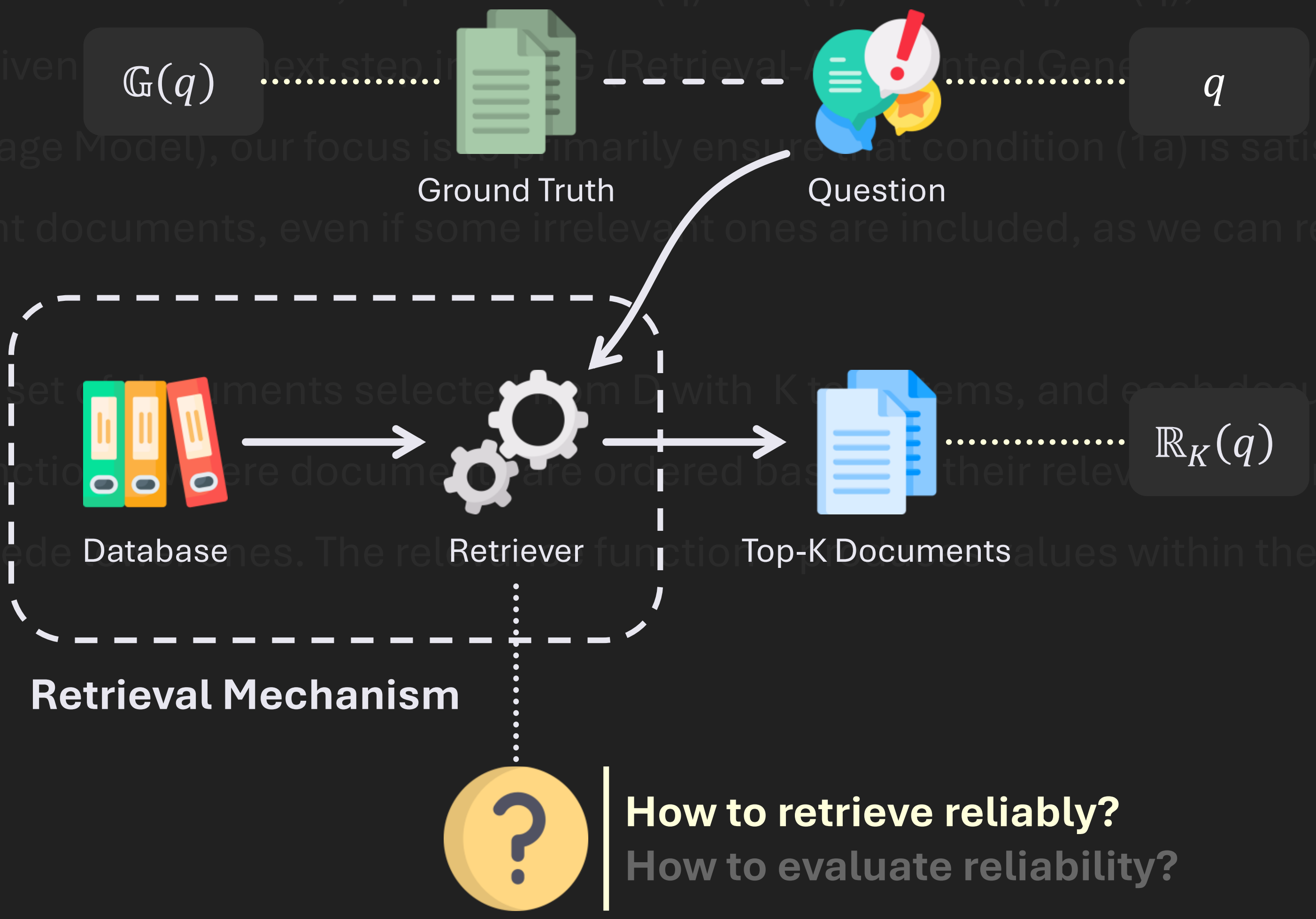
Methodology: Introduction

The first step in our methodology for this project is to define the primary task, which involves designing a retrieval mechanism to obtain an optimal set of documents, RK (q), that satisfies two key conditions: all relevant documents are retrieved while no irrelevant ones are included. These conditions, expressed as G(q)⊆RK (q) and RK (q)⊆G(q), ensure the precision of the retrieval process. However, given that the next step in a RAG (Retrieval-Augmented Generation) system is a response-generating LLM (Large Language Model), our focus is t[illegible]marily ensure th[illegible]condition (1a) is satisfied. This means we prioritize retrieving all relevant do[illegible] even if so[illegible]elevant ones[illegible]cluded, as[illegible]ely on the LLM's capability to handle the redundancy.

Thus, we define RK (q) as the set of documents selected from D with K total items, and each document's relevance is measured using a ranking function r, where documents are [illegible]ered based on their relevance score r(q,di), ensuring that higher relevance scores prece[illegible]nes. The relev[illegible] function r p[illegible]es values [illegible] range [0,1].



Methodology: Introduction



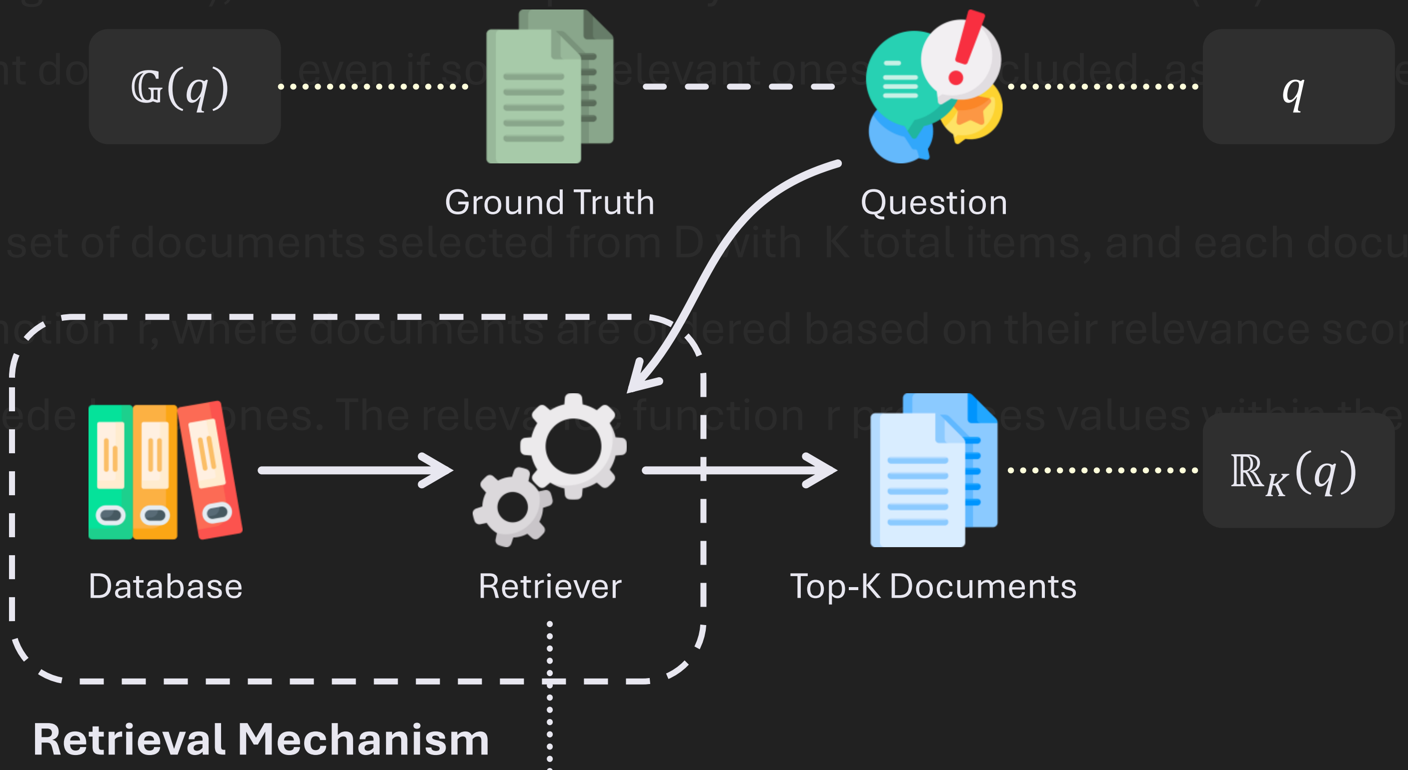
Methodology: Introduction

Do we have redundancy in what we retrieve?

$$\mathbb{R}_K(q) \subseteq \mathbb{G}(q)$$

Do we have all the relevant information?

$$\mathbb{G}(q) \subseteq \mathbb{R}_K(q)$$



How to retrieve reliably?

How to evaluate reliability?

ChatGT3

Explore GT3s

Today

- Project Objectives
- Problem Statement
- Quick Solution
- Methodology: Introduction
- Methodology: Metrics
- Methodology: Baselines
- Proposed Method
- Comparisons
- Conclusion

Enhanced Retrieval

Smarter Responses

ChatGT3

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What metrics were used and introduced to evaluate the retrieval mechanisms?

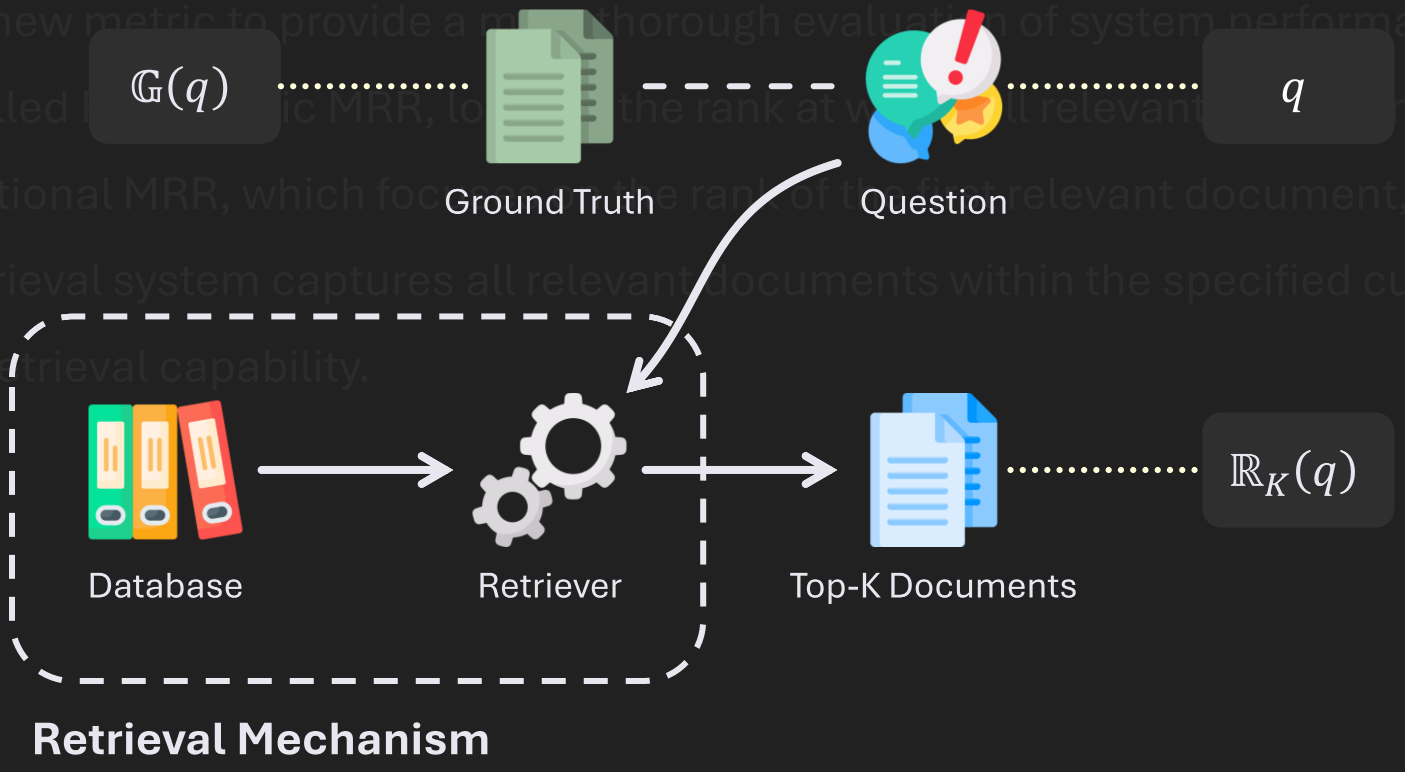
Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Methodology: Metrics

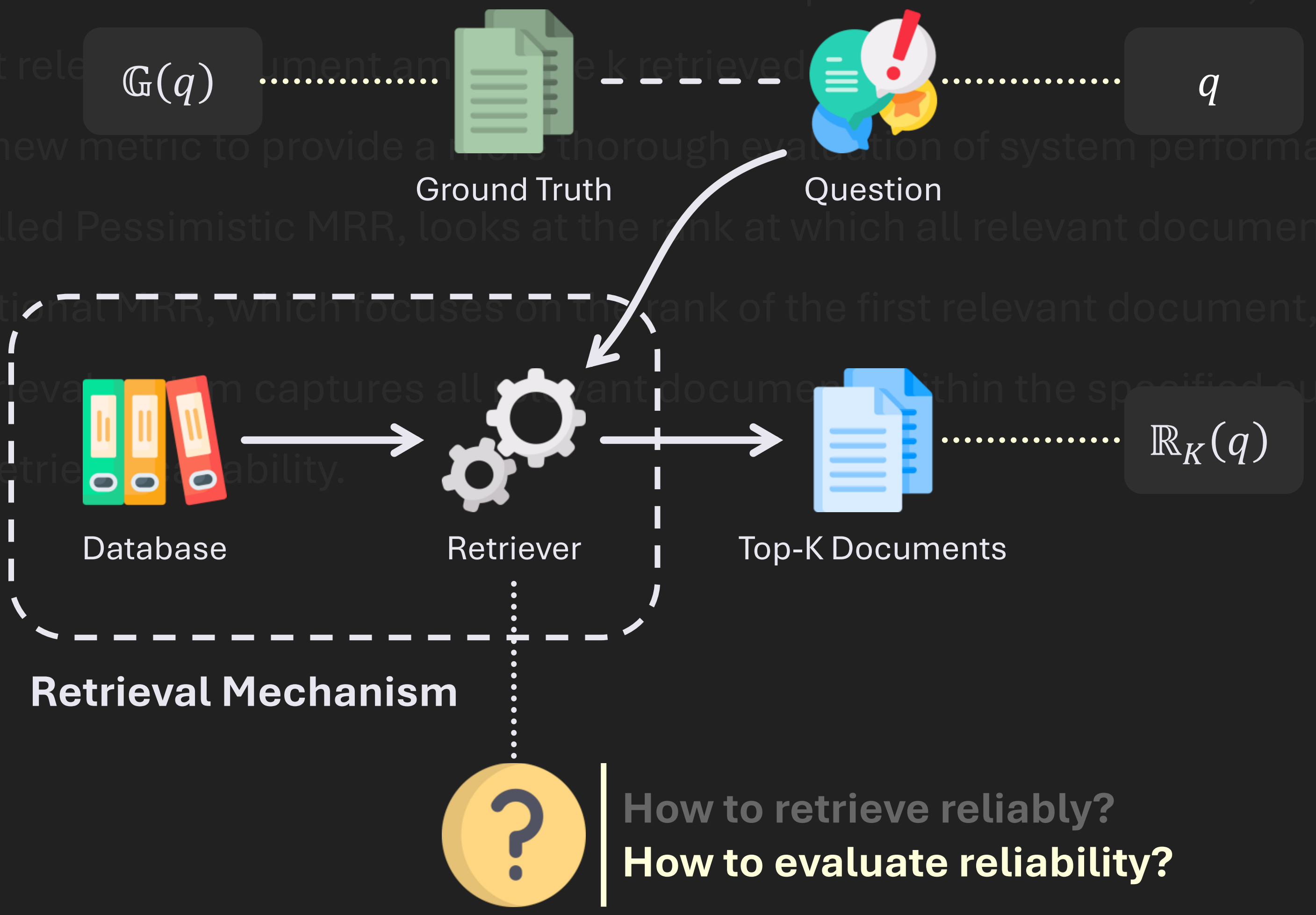
To evaluate the effectiveness of the methods used in this study, we rely on three key metrics. Let Q represent the set of all potential questions we want to assess. The first two metrics are widely used in information retrieval tasks. The first, Recall@k, calculates the proportion of relevant documents retrieved within the top k results. The second, Mean Reciprocal Rank (MRR), measures the rank of the first relevant document among the k retrieved ones.

Additionally, we introduce a new [illegible] thorough evalua[illegible] of system performance across different aspects. This new metric, called [illegible]ic MRR, to[illegible] the rank at w[illegible]ll relevant[illegible]nts are included within the top k results. Unlike the traditional MRR, which foc[illegible]e ran[illegible] relevant document, Pessimistic MRR emphasizes how well the retrieval system captures all relevant documents within the specified cutoff, offering a more comprehensive measure of r[illegible]trieval capability.



Methodology: Metrics

To evaluate the effectiveness of the methods used in this study, we rely on three key metrics. Let Q represent the set of all potential questions we want to assess. The first two metrics are widely used in information retrieval tasks. The first, Recall@k, calculates the proportion of relevant documents retrieved within the top k results. The second, Mean Reciprocal Rank (MRR), measures the rank of the first rele [illegible] ument am [illegible] e k retrieved [illegible] Additionally, we introduce a new metric to provide a [illegible] thorough ev [illegible] ation of system performance across different aspects. This new metric, called Pessimistic MRR, looks at the r [illegible] nk at which all relevant documents are included within the top k results. Unlike the traditi [illegible] nal MRR, which focuses on th [illegible] rank of the first relevant document, Pessimistic MRR emphasizes how well the retr [illegible] eva [illegible] m captures all [illegible] vant docume [illegible] ithin the sp [illegible] utoff, offering a more comprehensive measure of re [illegible] trie [illegible] ability.



Methodology: Metrics

$$R_K = \frac{|\mathbb{R}_K(q) \cap \mathbb{G}(q)|}{|\mathbb{G}(q)|}$$

Recall

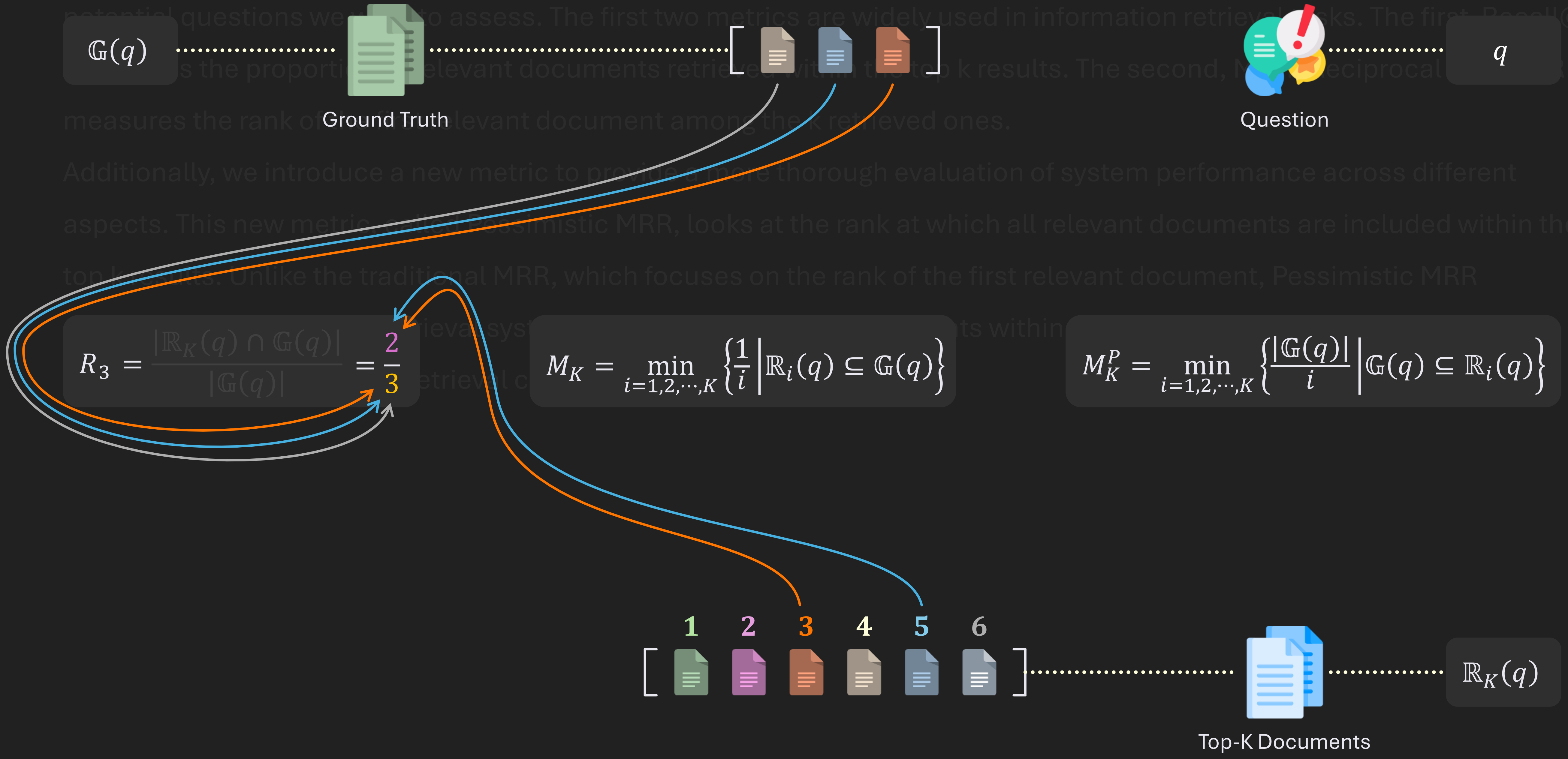
$$M_K = \min_{i=1,2,\cdots,K} \left\{ \frac{1}{i} \,\middle|\, \mathbb{R}_i(q) \subseteq \mathbb{G}(q) \right\}$$

Mean Reciprocal Rank (MRR)

$$M_K^P = \min_{i=1,2,\cdots,K} \left\{ \frac{|\mathbb{G}(q)|}{i} \,\middle|\, \mathbb{G}(q) \subseteq \mathbb{R}_i(q) \right\}$$

Pessimistic MRR

Methodology: Metrics



$$R_3 = \frac{|\mathbb{R}_K(q) \cap \mathbb{G}(q)|}{|\mathbb{G}(q)|} = \frac{2}{3}$$

$$M_K = \min_{i=1,2,\cdots,K} \left\{ \frac{1}{i} \,\middle|\, \mathbb{R}_i(q) \subseteq \mathbb{G}(q) \right\}$$

$$M_K^P = \min_{i=1,2,\cdots,K} \left\{ \frac{|\mathbb{G}(q)|}{i} \,\middle|\, \mathbb{G}(q) \subseteq \mathbb{R}_i(q) \right\}$$

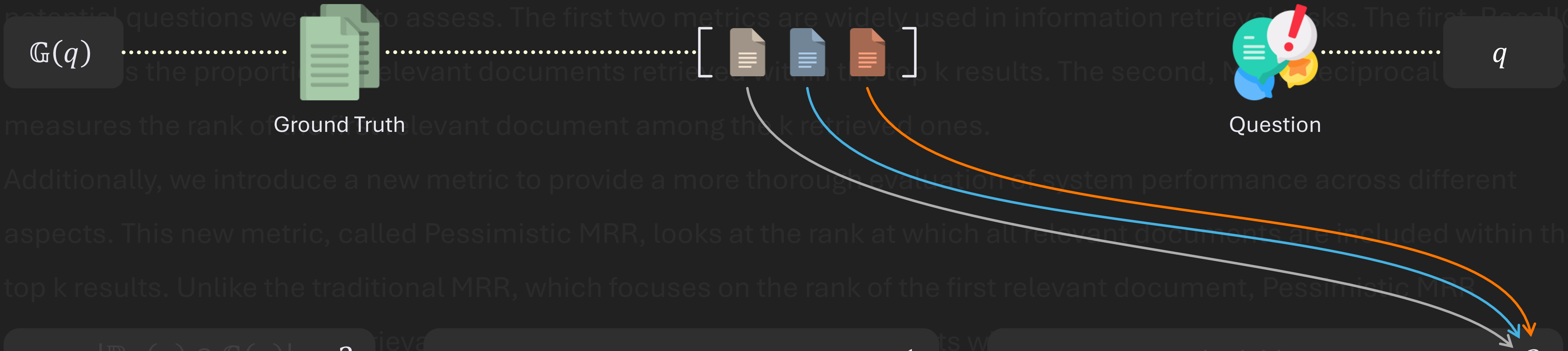
Methodology: Metrics

$$R_3 = \frac{|\mathbb{R}_K(q) \cap \mathbb{G}(q)|}{|\mathbb{G}(q)|} = \frac{2}{3}$$

$$M_4 = \min_{i=1,2,\cdots,K} \left\{ \frac{1}{i} \middle| \mathbb{R}_i(q) \subseteq \mathbb{G}(q) \right\} = \frac{1}{3}$$

$$M_K^P = \min_{i=1,2,\cdots,K} \left\{ \frac{|\mathbb{G}(q)|}{i} \middle| \mathbb{G}(q) \subseteq \mathbb{R}_i(q) \right\}$$

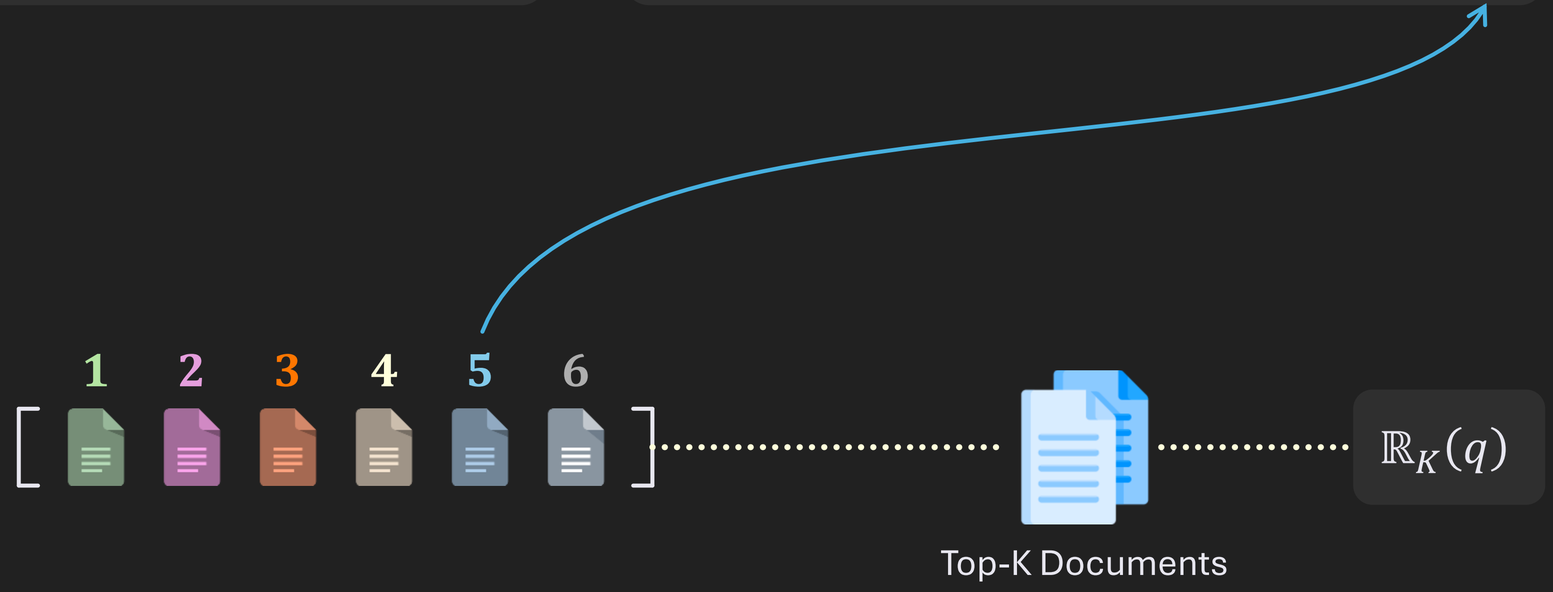
Methodology: Metrics



$$R_3 = \frac{|\mathbb{R}_K(q) \cap \mathbb{G}(q)|}{|\mathbb{G}(q)|} = \frac{2}{3}$$

$$M_4 = \min_{i=1,2,\cdots,K} \left\{ \frac{1}{i} \,\middle|\, \mathbb{R}_i(q) \subseteq \mathbb{G}(q) \right\} = \frac{1}{3}$$

$$M_5^P = \min_{i=1,2,\cdots,K} \left\{ \frac{|\mathbb{G}(q)|}{i} \,\middle|\, \mathbb{G}(q) \subseteq \mathbb{R}_i(q) \right\} = \frac{3}{5}$$



ChatGT3

Explore GT3s

Today

- Project Objectives 💎
- Problem Statement ❓
- Quick Solution 🌟
- Methodology: Introduction
- Methodology: Metrics
- Methodology: Baselines
- Proposed Method
- Comparisons
- Conclusion

Enhanced Retrieval
Smarter Responses

ChatGT3

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What are the baselines introduced and considered in this study?

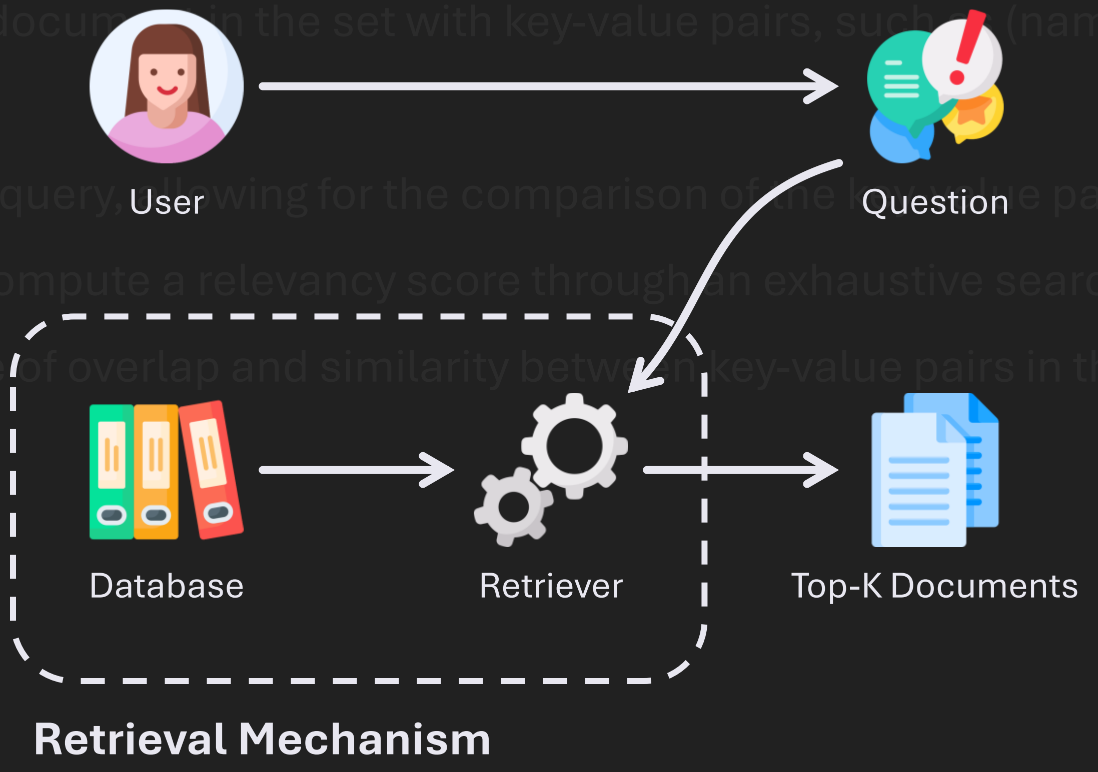
Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and topics as entity types, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each document in the set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

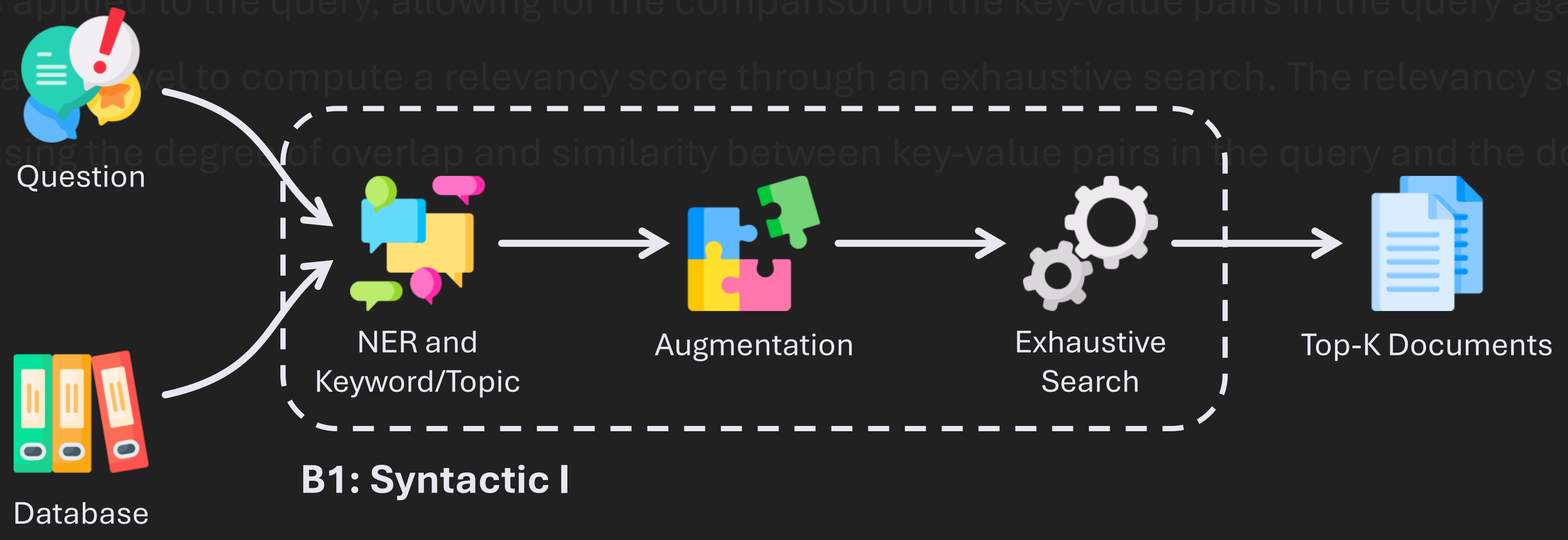
This same process is applied to the query, allowing for the comparison of the key-value pairs in the query against those in the documents at a character level to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.



Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and topics as entity types, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each document in the set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

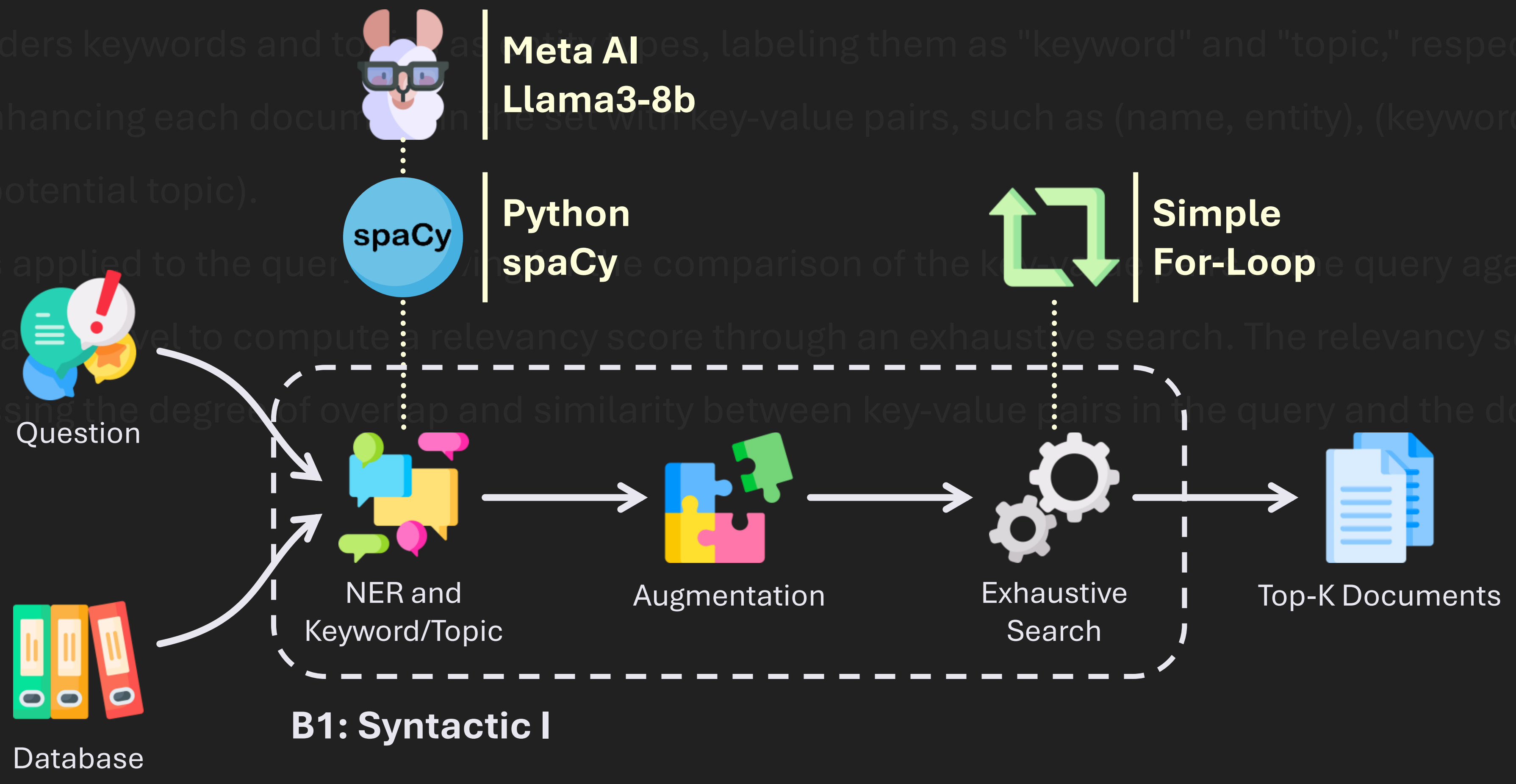
This same process is applied to the query, allowing for the comparison of the key-value pairs in the query against those in the documents at a character level to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.



Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and to [illegible] as [illegible] pes, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each docum [illegible] in [illegible] he set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

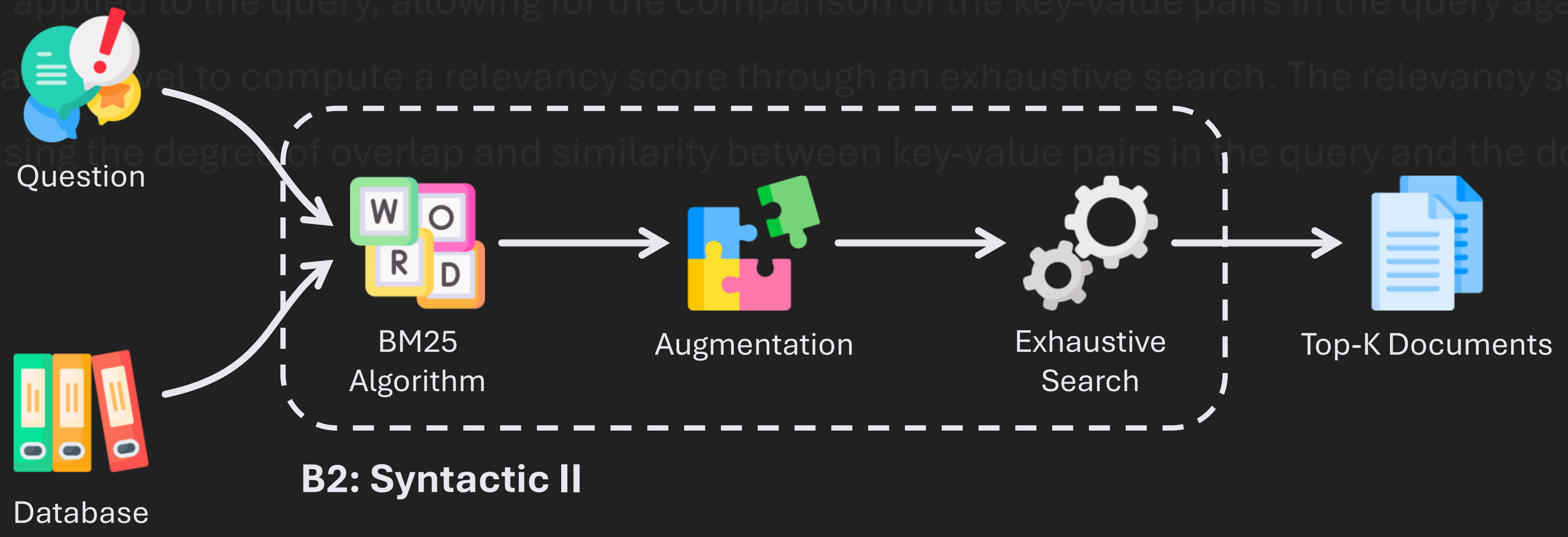
This same process is applied to the quer [illegible] comparison of the k [illegible] e query against those in the documents at a chara [illegible] el to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.



Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and topics as entity types, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each document in the set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

This same process is applied to the query, allowing for the comparison of the key-value pairs in the query against those in the documents at a character-level to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.

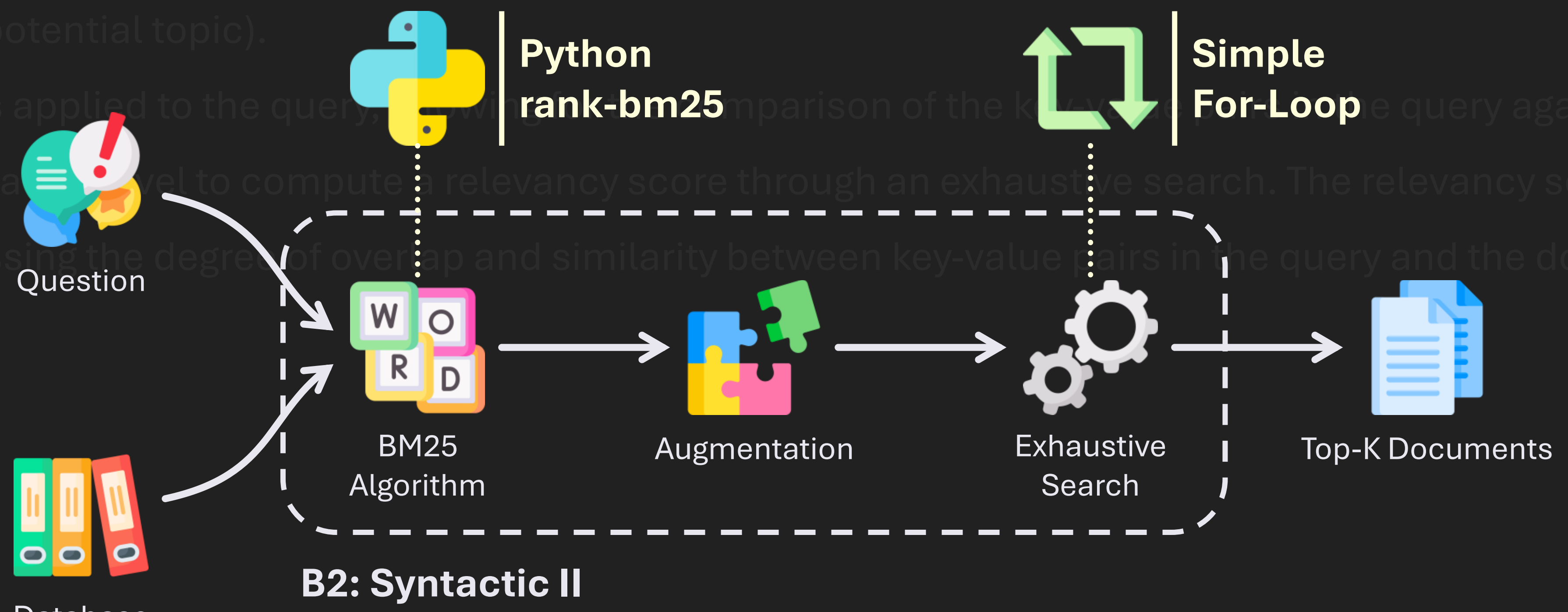


B2: Syntactic II

Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and topics as entity types, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each document in the set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

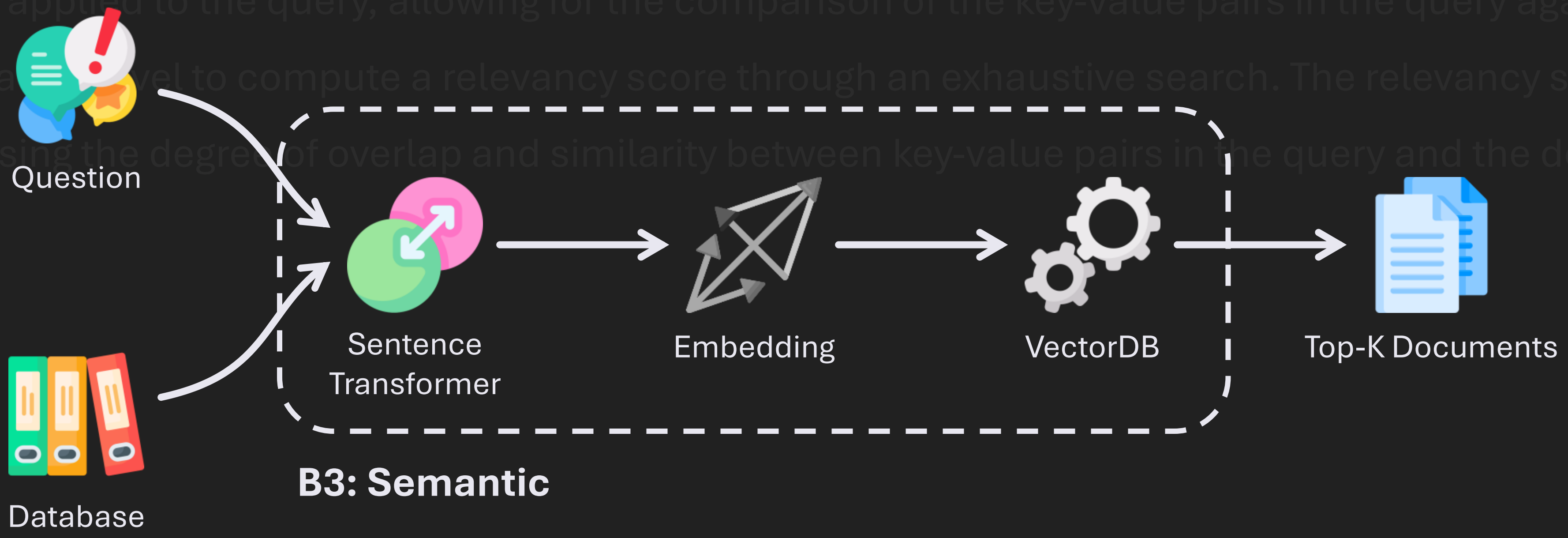
This same process is applied to the query, ... comparison of the k... e query against those in the documents at a chara... el to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.



Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and topics as entity types, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each document in the set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

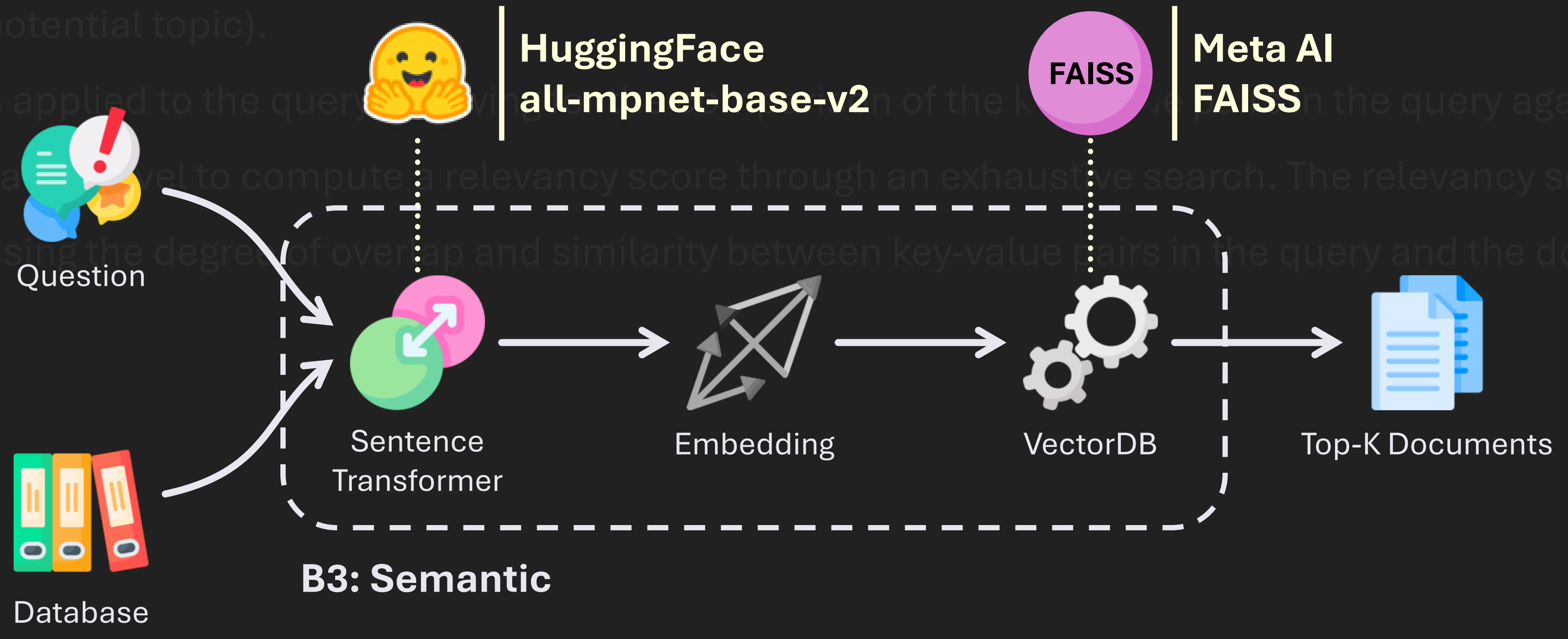
This same process is applied to the query, allowing for the comparison of the key-value pairs in the query against those in the documents at a character level to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.



Methodology: Baselines

The first baseline is Syntactic Search I (B1). This approach, combines Named Entity Recognition (NER) with keyword and topic extraction to perform a character-level search within a given document set for a specified query. B1 builds on the concept of key-value pairs derived from NER, extending it to include keywords and topics by treating them as additional entities. Essentially, B1 considers keywords and topics as entity types, labeling them as "keyword" and "topic," respectively. The process begins by enhancing each document in the set with key-value pairs, such as (name, entity), (keyword, potential keyword), or (topic, potential topic).

This same process is applied to the quer [illegible] of the k [illegible] the query against those in the documents at a chara [illegible] el to compute a relevancy score through an exhaustive search. The relevancy score is determined by assessing the degree of overlap and similarity between key-value pairs in the query and the documents.



ChatGT3

Explore GT3s

Today

- Project Objectives
- Problem Statement
- Quick Solution
- Methodology: Introduction
- Methodology: Metrics
- Methodology: Baselines
- Proposed Method
- Comparisons
- Conclusion

Enhanced Retrieval

Smarter Responses

ChatGT3

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What is the novel method introduced in this study?

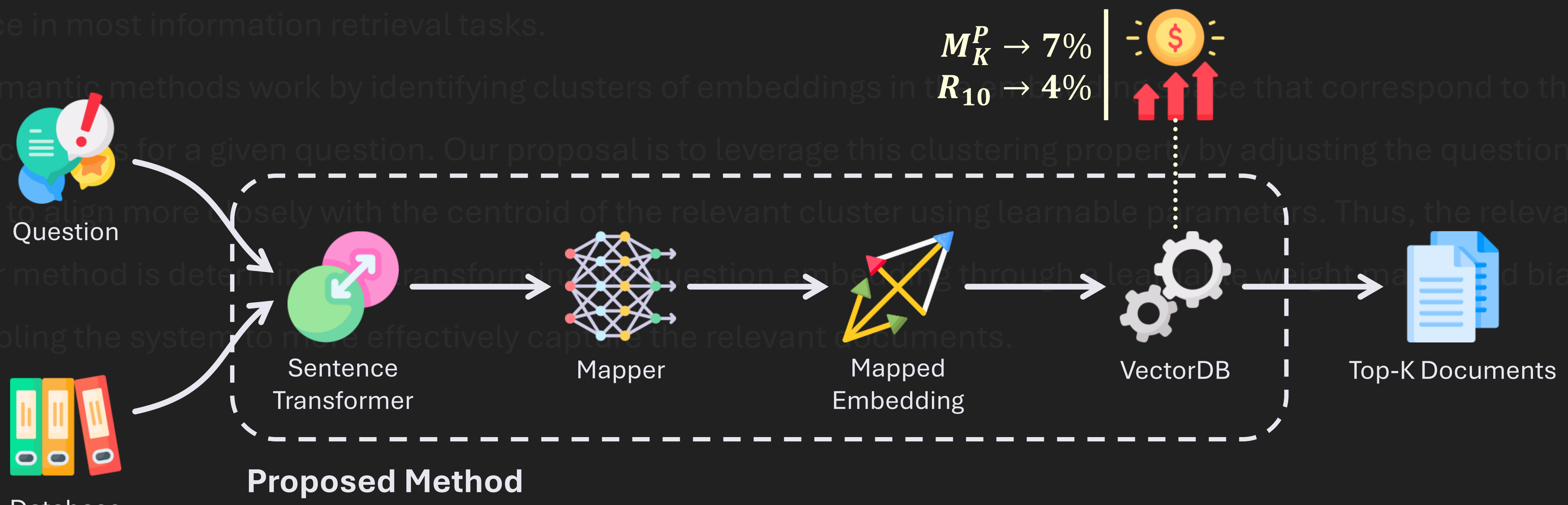
Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Proposed Approach

Once the baseline is established, the next step in our methodology is to develop an enhanced retrieval mechanism. Building on the semantic search approach from the baseline, our goal is to improve how the question embedding aligns with the relevant document embeddings. This approach is motivated by two key insights from the literature. First, semantic retrieval methods consistently outperform syntactic methods in capturing the subtle relationships between questions and documents, making it less useful to explore syntactic methods further due to their limitations in achieving optimal performance in most information retrieval tasks.

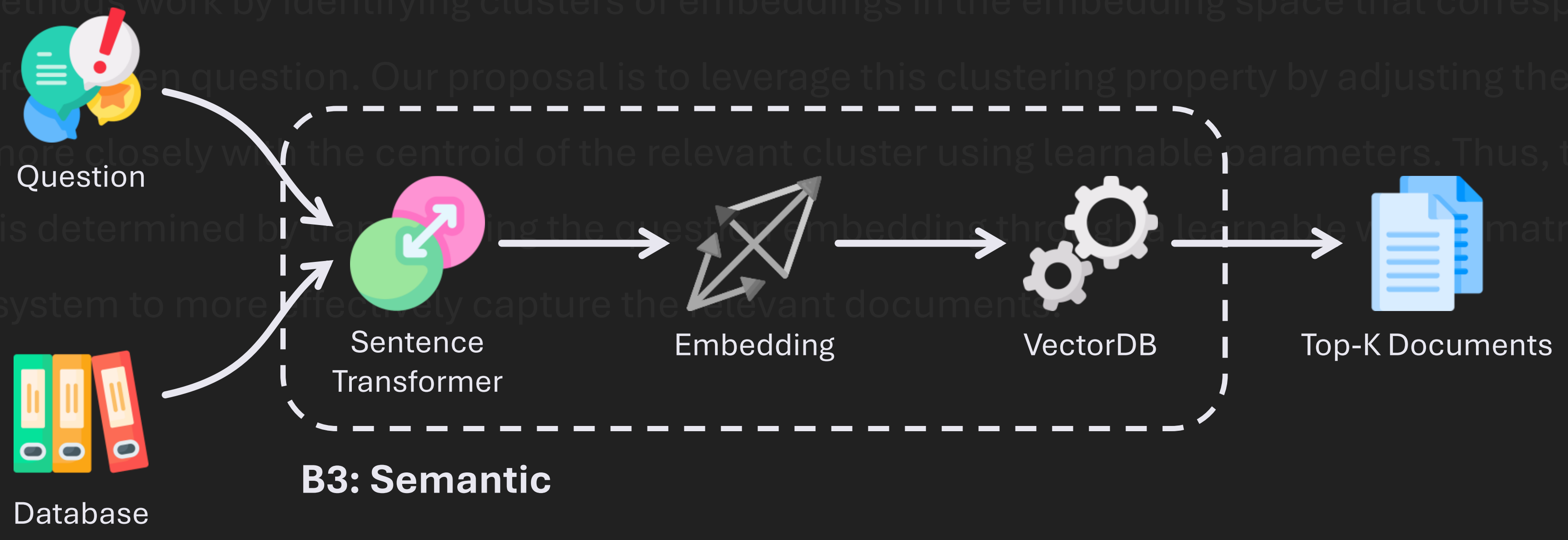
Second, semantic methods work by identifying clusters of embeddings in t[illegible]e that correspond to the most relevant doc[illegible]s for a given question. Our proposal is to leverage this clustering property by adjusting the question embedding to align more closely with the centroid of the relevant cluster using learnable parameters. Thus, the relevancy score in our method is dete[illegible]ransforming [illegible]uestion embe[illegible] through [illegible] weight ma[illegible]d bias vector, enabling the system to more effectively capture the relevant documents.



Proposed Approach

Once the baseline is established, the next step in our methodology is to develop an enhanced retrieval mechanism. Building on the semantic search approach from the baseline, our goal is to improve how the question embedding aligns with the relevant document embeddings. This approach is motivated by two key insights from the literature. First, semantic retrieval methods consistently outperform syntactic methods in capturing the subtle relationships between questions and documents, making it less useful to explore syntactic methods further due to their limitations in achieving optimal performance in most information retrieval tasks.

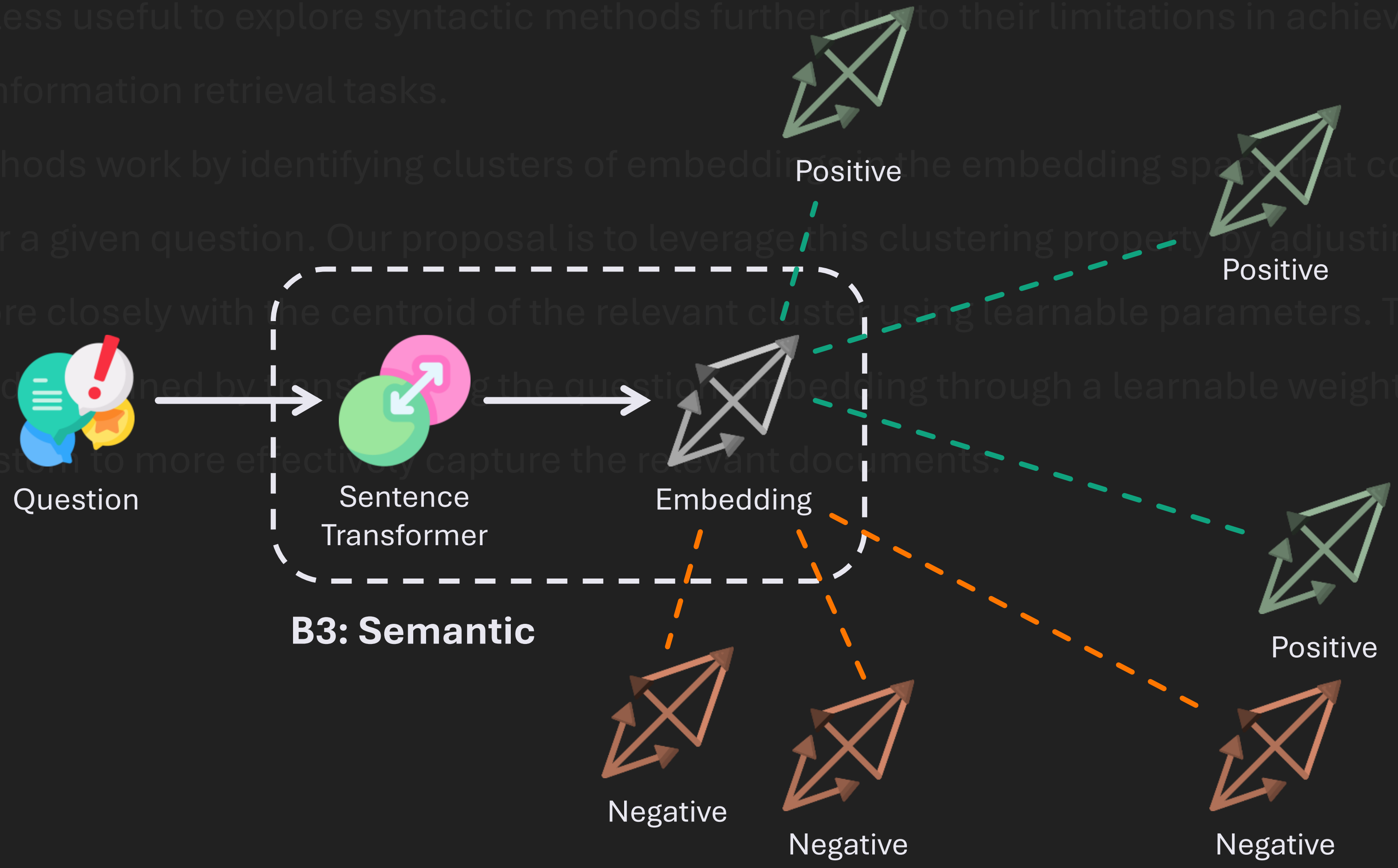
Second, semantic methods work by identifying clusters of embeddings in the embedding space that correspond to the most relevant documents f[illegible]en question. Our proposal is to leverage this clustering property by adjusting the question embedding to align more closely w[illegible]h the centroid of the relevant cluster using learnable parameters. Thus, the relevancy score in our method is determined b[illegible]ing the quest[illegible]mbedding thr[illegible]gh a learnable w[illegible]matrix and bias vector, enabling the system to more effectively capture the relevant documents.



Proposed Approach

Once the baseline is established, the next step in our methodology is to develop an enhanced retrieval mechanism. Building on the semantic search approach from the baseline, our goal is to improve how the question embedding aligns with the relevant document embeddings. This approach is motivated by two key insights from the literature. First, semantic retrieval methods consistently outperform syntactic methods in capturing the subtle relationships between questions and documents, making it less useful to explore syntactic methods further due to their limitations in achieving optimal performance in most information retrieval tasks.

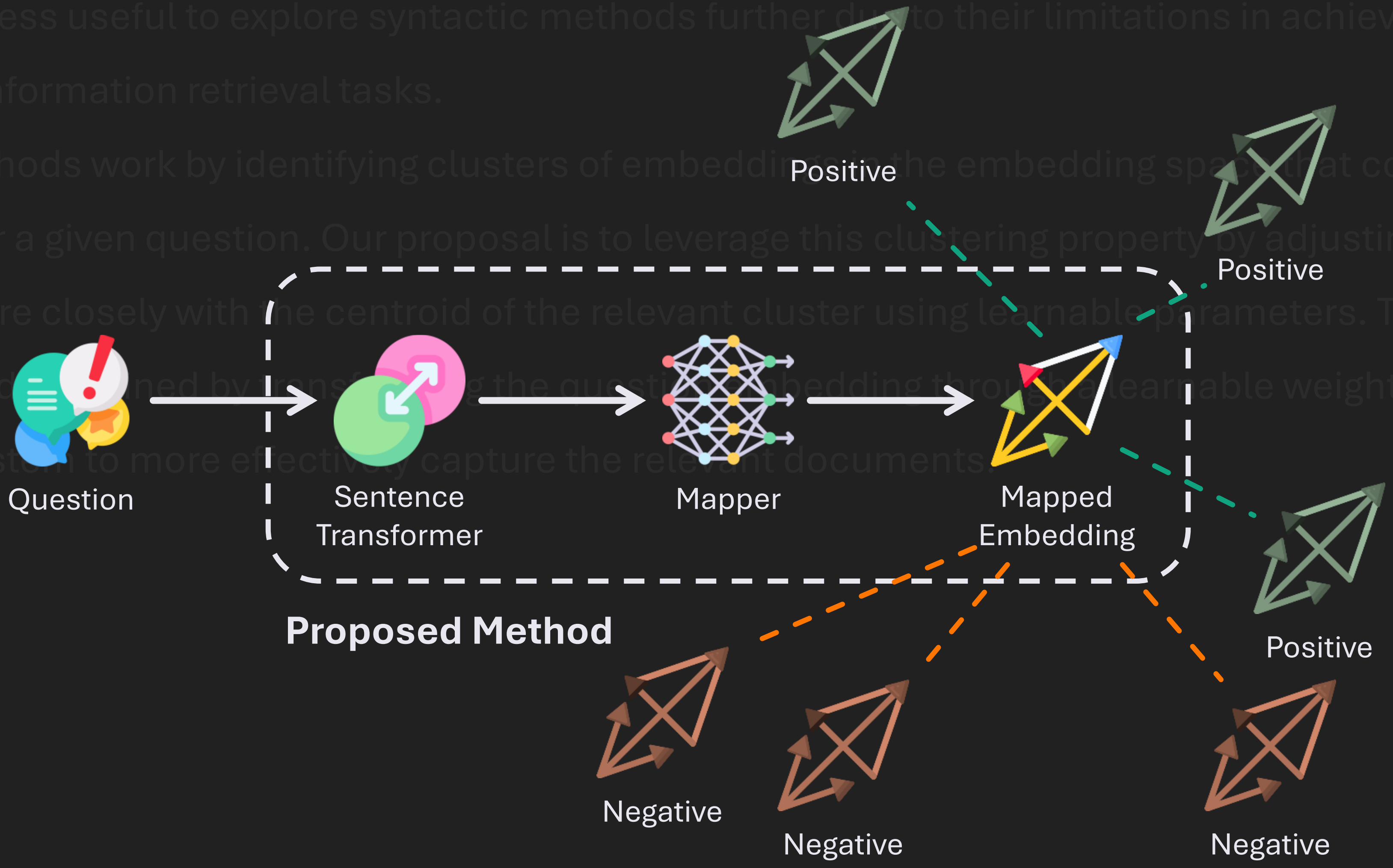
Second, semantic methods work by identifying clusters of embeddings in the embedding space that correspond to the most relevant documents for a given question. Our proposal is to leverage this clustering property by adjusting the question embedding to align more closely with the centroid of the relevant cluster using learnable parameters. Thus, the relevancy score in our method is determined by transforming the question embedding through a learnable weight matrix and bias vector, enabling the system to more effectively capture the relevant documents.



Proposed Approach

Once the baseline is established, the next step in our methodology is to develop an enhanced retrieval mechanism. Building on the semantic search approach from the baseline, our goal is to improve how the question embedding aligns with the relevant document embeddings. This approach is motivated by two key insights from the literature. First, semantic retrieval methods consistently outperform syntactic methods in capturing the subtle relationships between questions and documents, making it less useful to explore syntactic methods further due to their limitations in achieving optimal performance in most information retrieval tasks.

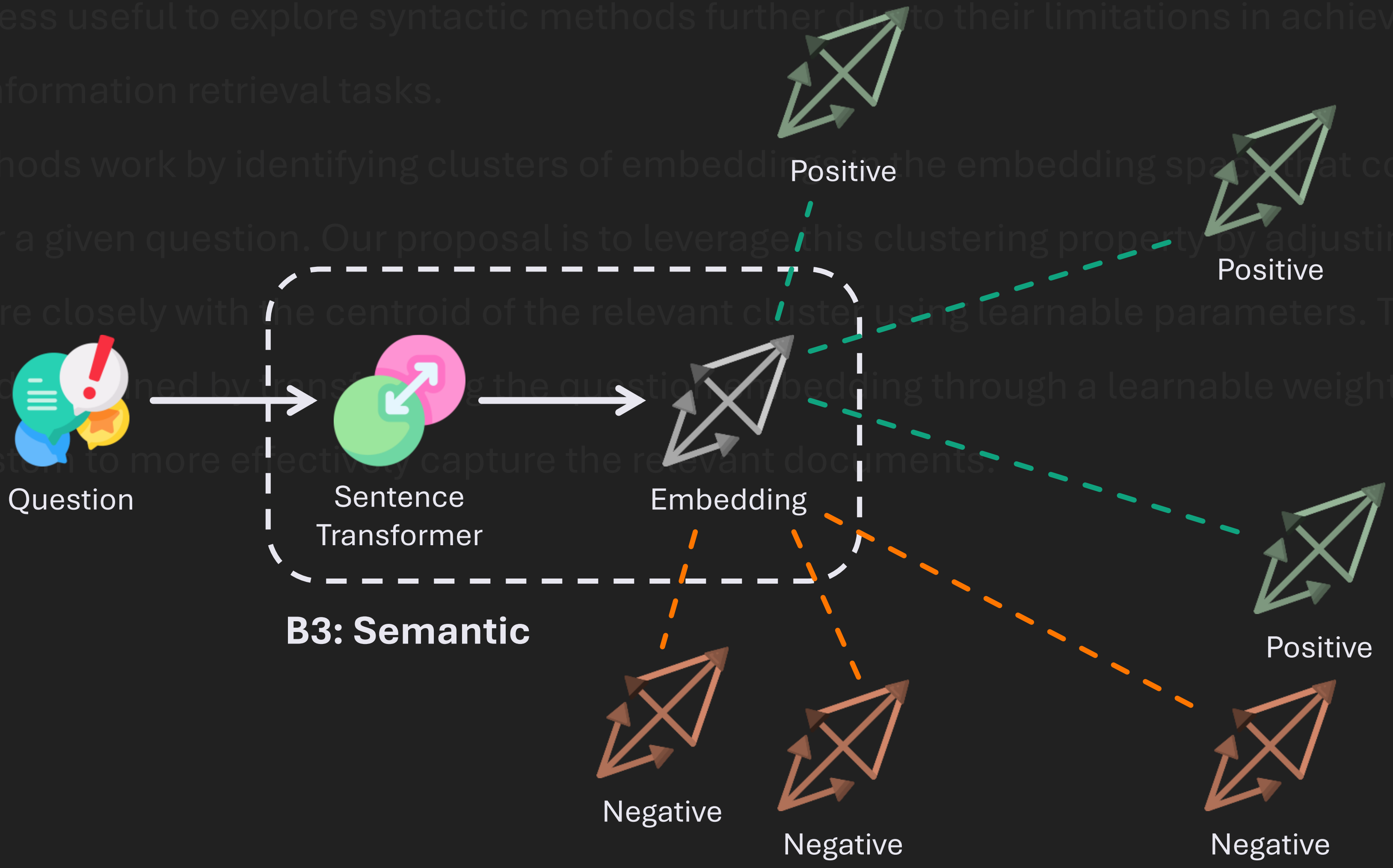
Second, semantic methods work by identifying clusters of embeddings in the embedding space that correspond to the most relevant documents for a given question. Our proposal is to leverage this clustering property by adjusting the question embedding to align more closely with the centroid of the relevant cluster using learnable parameters. Thus, the relevancy score in our method is determined by transforming the question embedding through a learnable weight matrix and bias vector, enabling the system to more effectively capture the relevant documents.



Proposed Approach

Once the baseline is established, the next step in our methodology is to develop an enhanced retrieval mechanism. Building on the semantic search approach from the baseline, our goal is to improve how the question embedding aligns with the relevant document embeddings. This approach is motivated by two key insights from the literature. First, semantic retrieval methods consistently outperform syntactic methods in capturing the subtle relationships between questions and documents, making it less useful to explore syntactic methods further due to their limitations in achieving optimal performance in most information retrieval tasks.

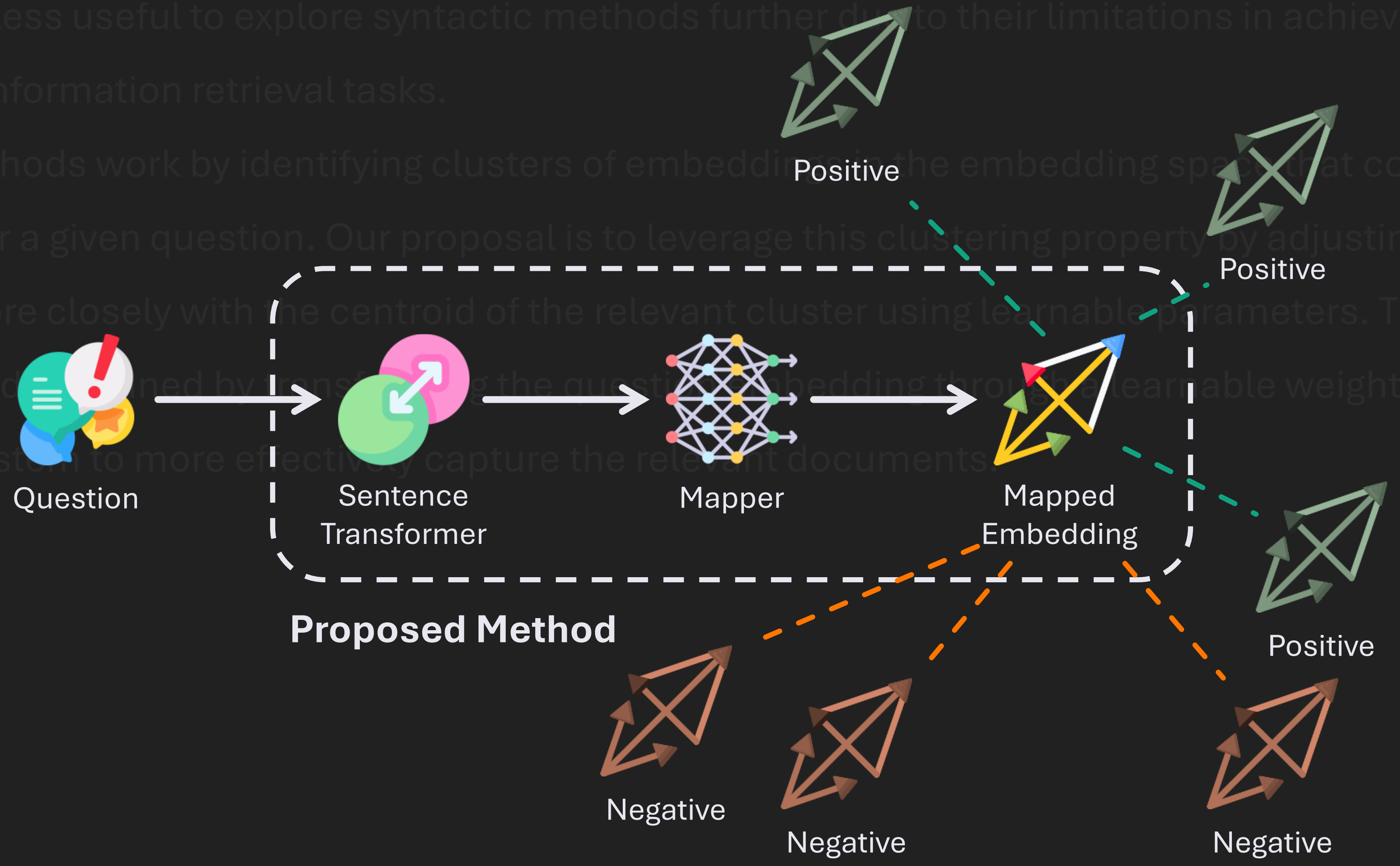
Second, semantic methods work by identifying clusters of embeddings in the embedding space that correspond to the most relevant documents for a given question. Our proposal is to leverage this clustering property by adjusting the question embedding to align more closely with the centroid of the relevant cluster using learnable parameters. Thus, the relevancy score in our method is determined by transforming the question embedding through a learnable weight matrix and bias vector, enabling the system to more effectively capture the relevant documents.



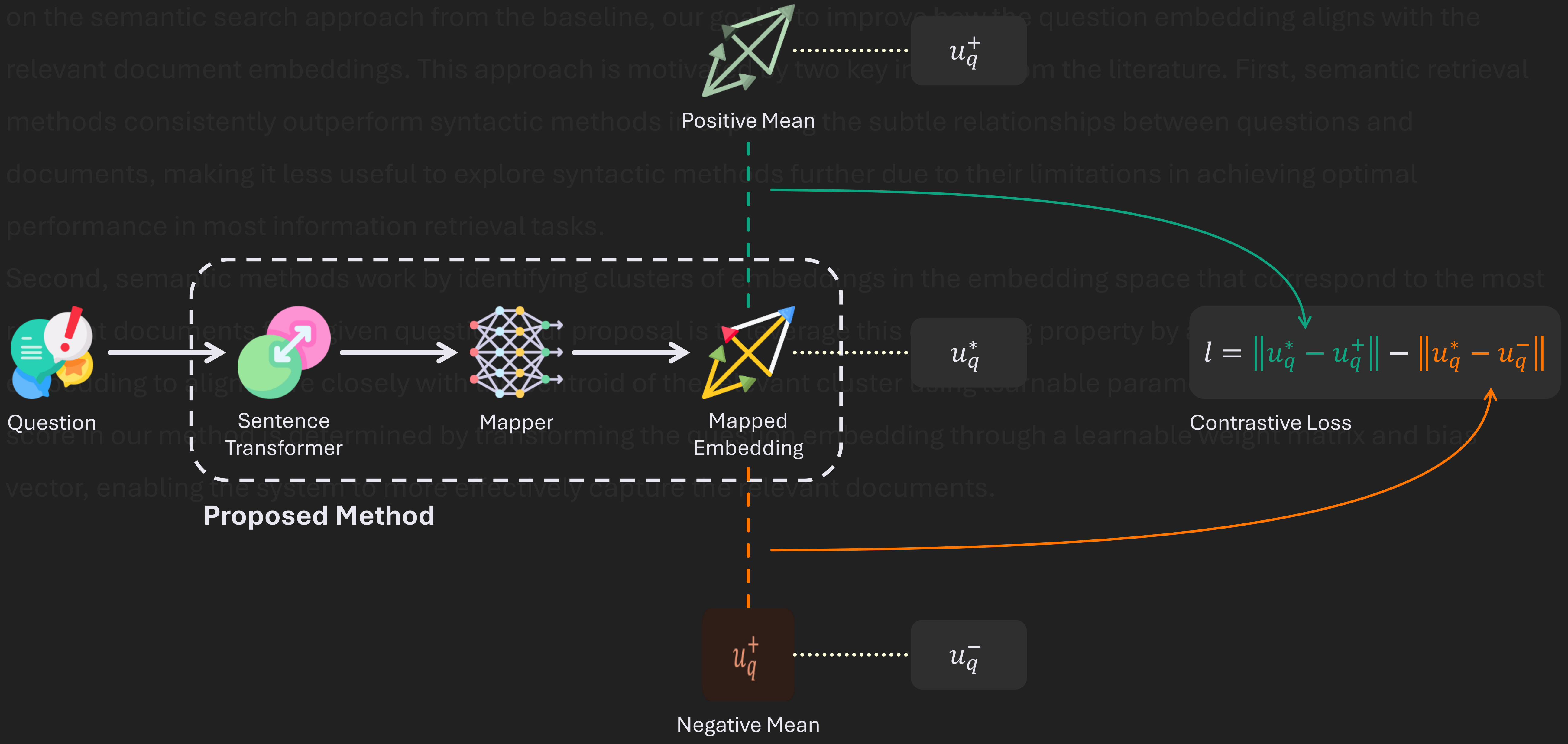
Proposed Approach

Once the baseline is established, the next step in our methodology is to develop an enhanced retrieval mechanism. Building on the semantic search approach from the baseline, our goal is to improve how the question embedding aligns with the relevant document embeddings. This approach is motivated by two key insights from the literature. First, semantic retrieval methods consistently outperform syntactic methods in capturing the subtle relationships between questions and documents, making it less useful to explore syntactic methods further due to their limitations in achieving optimal performance in most information retrieval tasks.

Second, semantic methods work by identifying clusters of embeddings in the embedding space that correspond to the most relevant documents for a given question. Our proposal is to leverage this clustering property by adjusting the question embedding to align more closely with the centroid of the relevant cluster using learnable parameters. Thus, the relevancy score in our method is determined by transforming the question embedding through a learnable weight matrix and bias vector, enabling the system to more effectively capture the relevant documents.



Proposed Approach



ChatGT3

ChatGT3

Explore GT3s

Today

- Project Objectives 💎
- Problem Statement ❓
- Quick Solution 🌟
- Methodology: Introduction 💭
- Methodology: Metrics
- Methodology: Baselines 🎯
- Proposed Method 💡
- Comparisons
- Conclusion

Enhanced Retrieval
Smarter Responses

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What improvement does the result achieved by this study?

Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Comparisons

The study began by setting the retrieval capacity (K) to 50, a stricter value than the typical 100, and optimized configurations for parameters such as distance (Jaro-Winkler), aggregation (maximization), tokenization (lemmatization), and encoder (all-mpnet-base-v2). The baseline results revealed that B3 outperformed others on the MS-Marco dataset, making it the primary baseline, while B2 was more effective on the Hotpot-QA dataset, highlighting the importance of dataset-specific tuning. Hyperparameter tuning was performed using Weights & Biases, exploring 1643 configurations to maximize MPK on a validation set from MS-Marco. Key findings included the optimal setting of ρ=0.75, where the mapper positioned question embeddings closer to relevant positives, and a preference for selecting the farthest positives and negatives as anchors, enhancing generalization by increasing the margin between positives and negatives, ultimately improving robustness and the clarity of boundaries in unseen data.

Dataset	Strategy	M_{50}	M_{50}^{P}	R_1	R_5	R_{10}
MS-Marco	B1	0.91	0.37	0.11	0.49	0.69
	B2	0.96	0.45	0.12	0.53	0.76
	B3	1.00	0.90	0.13	0.64	0.97
Hotpot-QA	B1	0.83	0.04	0.08	0.25	0.36
	B2	1.00	0.62	0.10	0.48	0.84
	B3	0.98	0.24	0.10	0.39	0.61

Comparisons

The study began by setting the retrieval capacity (K) to 50, a stricter value than the typical 100, and optimized configurations for parameters such as distance (Jaro-Winkler), aggregation (maximization), tokenization (lemmatization), and encoder (all-mpnet-base-v2). The baseline results revealed that B3 outperformed others on the MS-Marco dataset, making it the primary baseline, while B2 was more effective on the Hotpot-QA dataset, highlighting the importance of dataset-specific tuning.

Hyperparameter tuning was performed using Weights & Biases, exploring 1643 configurations to maximize MPK on a validation set from MS-Marco. Key findings included the optimal setting of ρ=0.75, where the mapper positioned question embeddings closer to relevant positives, and a preference for selecting the farthest positives and negatives as anchors, enhancing generalization by increasing the margin between positives and negatives, ultimately improving robustness and the clarity of boundaries in unseen data.

Dataset	Strategy	M_{50}	M^{P}_{50}	R_1	R_5	R_{10}
MS-Marco	B1	0.91	0.37	0.11	0.49	0.69
	B2	0.96	0.45	0.12	0.53	0.76
	B3	**1.00**	**0.90**	**0.13**	**0.64**	**0.97**
Hotpot-QA	B1	0.83	0.04	0.08	0.25	0.36
	B2	**1.00**	**0.62**	**0.10**	**0.48**	**0.84**
	B3	0.98	0.24	0.10	0.39	0.61

Comparisons

The study began by setting the retrieval capacity (K) to 50, a stricter value than the typical 100, and optimized configurations for parameters such as distance (Jaro-Winkler), aggregation (maximization), tokenization (lemmatization), and encoder (all-mpnet-base-v2). The baseline results revealed that B3 outperformed others on the MS-Marco dataset, making it the primary baseline, while B2 was more effective on the Hotpot-QA dataset, highlighting the importance of dataset-specific tuning. Hyperparameter tuning was performed using Weights & Biases, exploring 1643 configurations to maximize MPK on a validation set from MS-Marco. Key findings included the optimal setting of ρ=0.75, where the mapper positioned question embeddings closer to relevant positives, and a preference for selecting the farthest positives and negatives as anchors, enhancing generalization by increasing the margin between positives and negatives, ultimately improving robustness and the clarity of boundaries in unseen data.

Dataset	Strategy	M_{50}	M_{50}^{P}	R_1	R_5	R_{10}
MS-Marco	B3	0.98	0.77	0.12	0.61	0.91
	Mapper	0.99	0.84	0.13	0.61	0.95
Hotpot-QA	B3	0.95	0.12	0.10	0.31	0.43
	Mapper	0.82	0.10	0.08	0.26	0.38

Comparisons

The study began by setting the retrieval capacity (K) to 50, a stricter value than the typical 100, and optimized configurations for parameters such as distance (Jaro-Winkler), aggregation (maximization), tokenization (lemmatization), and encoder (all-mpnet-base-v2). The baseline results revealed that B3 outperformed others on the MS-Marco dataset, making it the primary baseline, while B2 was more effective on the Hotpot-QA dataset, highlighting the importance of dataset-specific tuning. Hyperparameter tuning was performed using Weights & Biases, exploring 1643 configurations to maximize MPK on a validation set from MS-Marco. Key findings included the optimal setting of ρ=0.75, where the mapper positioned question embeddings closer to relevant positives, and a preference for selecting the farthest positives and negatives as anchors, enhancing generalization by increasing the margin between positives and negatives, ultimately improving robustness and the clarity of boundaries in unseen data.

Dataset	Strategy	M_{50}	M^P_{50}	R_1	R_5	R_{10}
MS-Marco	B3	0.98	0.77	0.12	0.61	0.91
	Mapper	**0.99**	**0.84**	**0.13**	**0.61**	**0.95**
Hotpot-QA	**B3**	**0.95**	**0.12**	**0.10**	**0.31**	**0.43**
	Mapper	0.82	0.10	0.08	0.26	0.38

ChatGT3

ChatGT3

Explore GT3s

Today

- Project Objectives
- Problem Statement
- Quick Solution
- Methodology: Introduction
- Methodology: Metrics
- Methodology: Baselines
- Proposed Method
- Comparisons
- Conclusion

Enhanced Retrieval

Smarter Responses

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

What is the overall conclusion of this study?

Create Image | Code | Summarize | Get advice | More

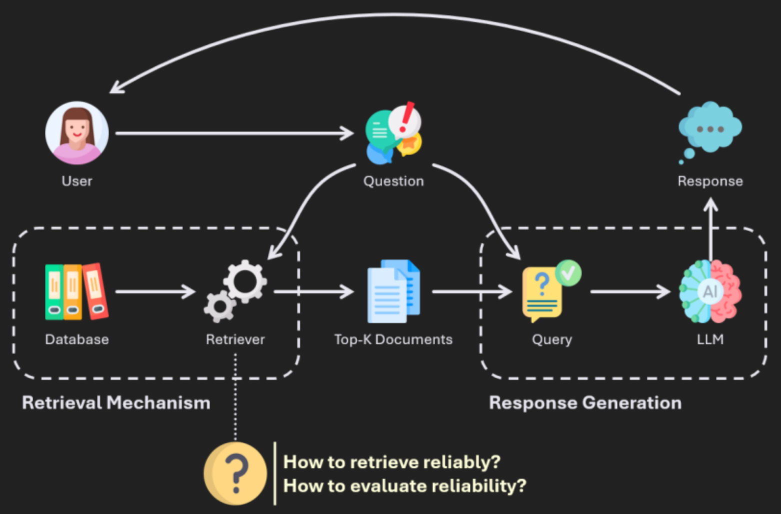
ChatGT3 can make mistakes. Check important info.

Conclusion

- **TWO QUESTIONS** were posed regarding the retrieval mechanism in the RAG architecture:
  - **How to Retrieve Reliably?**
  - **How to Measure Reliability?**
- **ANSWERS** to these question are provided by:
  - **MAPPER:** A Novel Method for **Enhanced Reliability** to **Achieve Smarter Responses**
  - **PESSIMISTIC MRR:** A Novel Metric to **Evaluate the Proposed Method**
- **RESULTS** can be summarized as follows:
  - **MAPPER OUTPERFORMS** successfully **all the baselines**.
  - **SEMANTIC RETRIEVAL** is not a **panacea**.

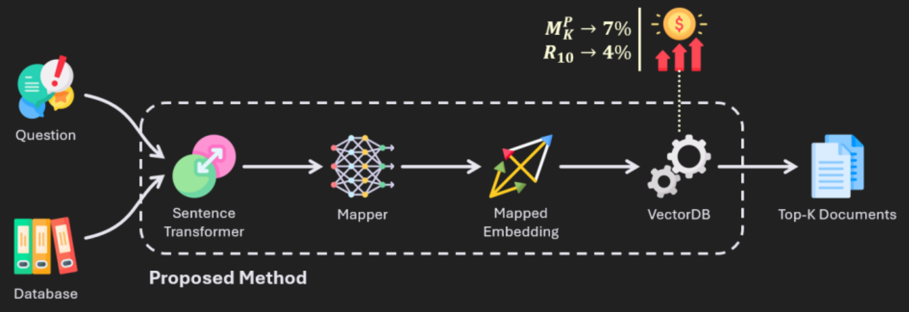
Conclusion

- TWO QUESTIONS were posed regarding the retrieval mechanism in the RAG architecture:
  - How to Retrieve Reliably?
  - How to Measure Reliability?
- ANSWERS to these question are provided by:
  - **MAPPER:** A Novel Method for **Enhanced Reliability** to **Achieve Smarter Responses**
  - **PESSIMISTIC MRR:** A Novel Metric to **Evaluate the Proposed Method**
- RESULTS can be summarized as follows:
  - **MAPPER OUTPERFORMS** successfully **all the baselines**.
  - **SEMANTIC RETRIEVAL** is not a **panacea**.



Conclusion

- **TWO QUESTIONS** were posed regarding the retrieval mechanism in the RAG architecture:
  - **How to Retrieve Reliably?**
  - **How to Measure Reliability?**
- **ANSWERS** to these question are provided by:
  - **MAPPER:** A Novel Method for **Enhanced Reliability** to **Achieve Smarter Responses**
  - **PESSIMISTIC MRR:** A Novel Metric to **Evaluate the Proposed Method**
- **RESULTS** can be summarized as follows:
  - **MAPPER OUTPERFORMS** successfully **all the baselines**.
  - **SEMANTIC RETRIEVAL** is not a **panacea**.



Conclusion

- **TWO QUESTIONS** were posed regarding the retrieval mechanism in the RAG architecture:
  - **How to Retrieve Reliably?**
  - **How to Measure Reliability?**
- **ANSWERS** to these question are provided by:
  - **MAPPER:** A Novel Method for **Enhanced Reliability** to **Achieve Smarter Responses**
  - **PESSIMISTIC MRR:** A Novel Metric to **Evaluate the Proposed Method**
- **RESULTS** can be summarized as follows:
  - **MAPPER OUTPERFORMS** successfully **all the baselines**.
  - **SEMANTIC RETRIEVAL** is not a **panacea**.

Dataset	Strategy	M_{50}	M^{P}_{50}	R_1	R_5	R_{10}
MS-Marco	B3	0.98	0.77	0.12	0.61	0.91
	Mapper	**0.99**	**0.84**	**0.13**	**0.61**	**0.95**
Hotpot-QA	**B3**	**0.95**	**0.12**	**0.10**	**0.31**	**0.43**
	Mapper	0.82	0.10	0.08	0.26	0.38

ChatGT3

Explore GT3s

Today

- Project Objectives 💎
- Problem Statement ❓
- Quick Solution 🌟
- Methodology: Introduction 💭
- Methodology: Metrics
- Methodology: Baselines 🎯
- Proposed Method 💡
- Comparisons
- Conclusion 🔨

Enhanced Retrieval

Smarter Responses

ChatGT3

What can I help with?

ADSP - P9 - RAG MARCO.pdf
PDF

T3 - REPORT .pdf
PDF

Is there any online demonstration for this project?

Create Image | Code | Summarize | Get advice | More

ChatGT3 can make mistakes. Check important info.

Demo

Enhanced Retrieval

Smarter Responses