

Oops!

Welcome back! Our systems are currently free, and we're excited to continue to develop retrieval part of the RAG. Please relax and enjoy the content.



Homayoun Afshari



Arash Daneshvar



Hossein Khodadadi

You had hit the free plan limit for GT3.

You can use GT3 or upgrade to GT3-4o to continue this chat.
Your limit is lifted since it's December 4th.

Continue

Get plus



ChatGT3



Explore GT3s

November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Gantt



Conclusion



Upgrade plan

What can I help with?



Message ChatGT3



Create Image



Code



Summarize



Get advice

More





ChatGT3



Explore GT3s

November

Project definition ...

Project objectives ...

Project relevance to SDGs ...

Design: user personas ...

Design: customer journey ...

Design: functional diagram ...

Manage: WBS ...

Manage: Gantt ...

Conclusion ...

Search chats...



New chat

November



Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Upgrade plan





ChatGT3



Explore GT3s

November

- Project definition ...
- Project objectives ...
- Project relevance to SDGs ...
- Design: user personas ...
- Design: customer journey ...
- Design: functional diagram ...
- Manage: WBS ...
- Manage: Gantt ...
- Conclusion ...



Upgrade plan

objectives



New chat



November



Project objectives



Design: functional diagram



Conclusion





ChatGT3



Explore GT3s

Today

Project objectives



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Gantt



Conclusion



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

Please describe the objectives of the project explained in this file.



Let's dive into optimizing information quality in a RAG system, with a spotlight on retrieval and response generation.

When it comes down to it, the quality of information is everything.

Based on insights from the PDF, here's a breakdown of the main objectives:

- **Objective 1:** Enhance the Retrieval Process Using LLMs
- **Objective 2:** Elevate the Quality of Context for the Response Generator

And here's a potential **research question** to dig deeper:

“How optimizing the retrieval stage in a RAG system ensures high-quality and contextually accurate responses?”

Also, for this project, the **value proposition** could be:

“Boosting Accuracy through Advanced Retrieval”

This approach keeps the focus sharp - enhancing accurate retrieval and delivering relevant.



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Gantt



Conclusion



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

Please describe the objectives of the project explained in this file.



Let's dive into optimizing information quality in a RAG system, with a spotlight on retrieval and response generation.

When it comes down to it, the quality of information is everything.

Based on insights from the PDF, here's a breakdown of the main objectives:

- **Objective 1:** Enhance the Retrieval Process Using LLMs
- **Objective 2:** Elevate the Quality of Context for the Response Generator

And here's a potential **research question** to dig deeper:

“How optimizing the retrieval stage in a RAG system ensures high-quality and contextually accurate responses?”

Also, for this project, the **value proposition** could be:

“Boosting Accuracy through Advanced Retrieval”

This approach keeps the focus sharp - enhancing accurate retrieval and delivering relevant.



Great! Could you provide us with the updated objective for the project? |





Today

Project objectives

November

Project definition

Project objectives

Project relevance to SDGs

Design: user personas

Design: customer journey

Design: functional diagram

Manage: WBS

Manage: Gantt

Conclusion



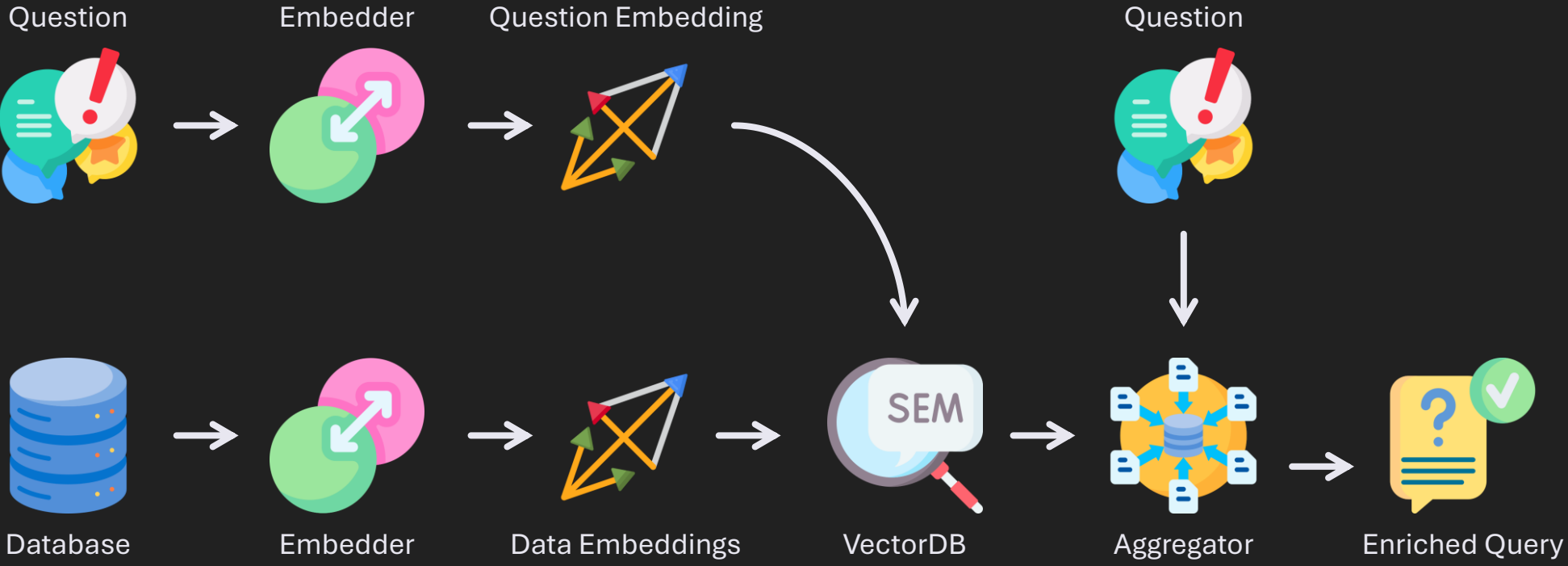
ADSP - P9 - RAG MARCO.pdf
PDF

Great! Could you provide us with the updated objective for the project?



Certainly! Building on the previous objectives, here is a refined breakdown of the updated objectives of the project:

- **Objective 1:** Refine the retrieval system.
- **Objective 2:** Improve the contextual quality of information fed into the response generation stage.



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Gantt



Conclusion



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

Great! Could you provide us with the updated objective for the project?

Search chats...



New chat

November



Project definition



Project objectives



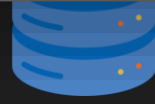
Project relevance to SDGs



Design: user personas



Design: customer journey



Database



Embedder



Data Embeddings



VectorDB



Aggregator



Enriched Query



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives

November

Project definition

Project objectives

Project relevance to SDGs

Design: user personas

Design: customer journey

Design: functional diagram

Manage: WBS

Manage: Gantt

Conclusion



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

Great! Could you provide us with the updated objective for the project?

research question

New chat

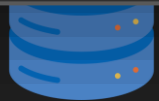
November

Project objectives

Design: customer journey

Manage: WBS

Conclusion



Database



Embedder



Data Embeddings



VectorDB



Aggregator



Enriched Query



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



Research question



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Cont...



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

Please describe the objectives of the project explained in this file.



Let's dive into optimizing information quality in a RAG system, with a spotlight on retrieval and response generation.

When it comes down to it, the quality of information is everything.

Based on insights from the PDF, here's a breakdown of the main objectives:

- **Objective 1:** Enhance the Retrieval Process Using LLMs
- **Objective 2:** Elevate the Quality of Context for the Response Generator

And here's a potential **research question** to dig deeper:

“How optimizing the retrieval stage in a RAG system ensures high-quality and contextually accurate responses?”

Also, for this project, the **value proposition** could be:

“Boosting Accuracy through Advanced Retrieval”

This approach keeps the focus sharp - enhancing accurate retrieval and delivering relevant.



Message ChatGT3





Today

Project objectives



Research question



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Cont...



ADSP - P9 - RAG MARCO.pdf
PDF

Please describe the objectives of the project explained in this file.



Let's dive into optimizing information quality in a RAG system, with a spotlight on retrieval and response generation.

When it comes down to it, the quality of information is everything.

Based on insights from the PDF, here's a breakdown of the main objectives:

- **Objective 1:** Enhance the Retrieval Process Using LLMs
- **Objective 2:** Elevate the Quality of Context for the Response Generator

And here's a potential **research question** to dig deeper:

“How optimizing the retrieval stage in a RAG system ensures high-quality and contextually accurate responses?”

Also, for this project, the **value proposition** could be:

“Boosting Accuracy through Advanced Retrieval”

This approach keeps the focus sharp - enhancing accurate retrieval and delivering relevant.



Great! please provide an updated version with a focus on the new direction. |





Today

Project objectives

Research question

November

Project definition

Project objectives

Project relevance to SDGs

Design: user personas

Design: customer journey

Design: functional diagram

Manage: WBS

Manage: Contt



ADSP - P9 - RAG MARCO.pdf
PDF

Great! please provide an updated version with a focus on the new direction.

Now you can view two options: text-based and visualized responses.



Response 1

Here's a potential **research question** to dig deeper:

“How optimizing the retrieval stage in a RAG system ensures high-quality and contextually accurate responses?”

And the **value proposition**:

“Enhanced Retrieval, Delivering Smarter Responses”

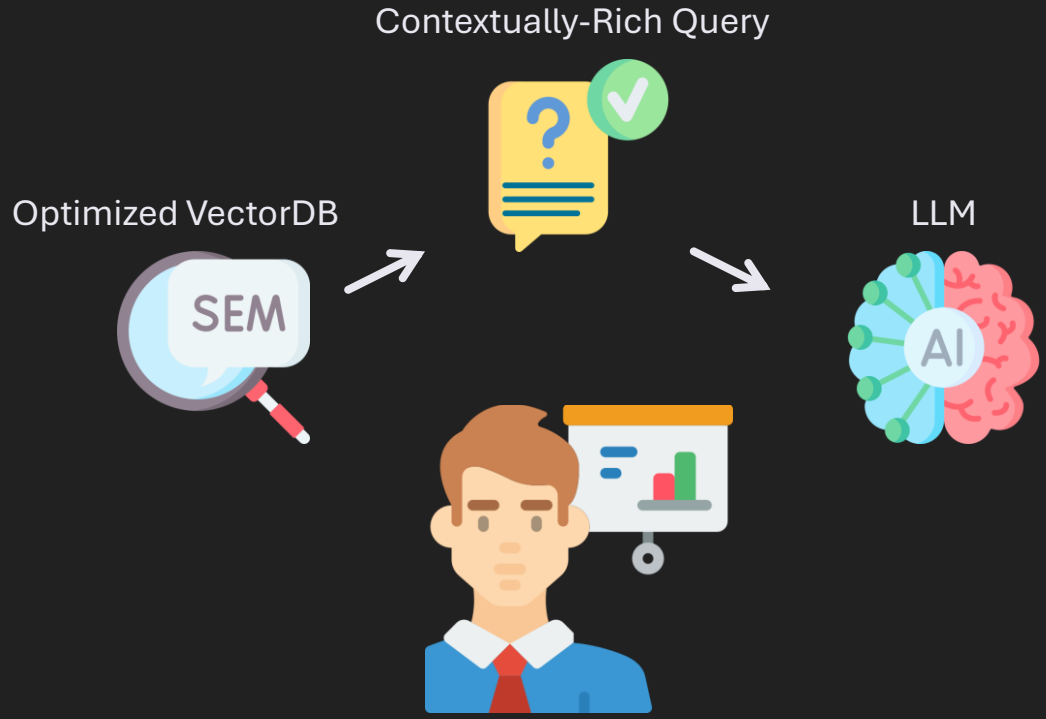
Aligned with the following user persona:

“Michael Foster, the AI software manager”



Response 2

The visualization is as follows:



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



Research question



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage: WBS



Manage: Cont...



Upgrade plan

What can I help with?



ADSP - P9 - RAG MARCO.pdf
PDF



Now, let's move to development. What is the main task a how to set a baseline? |



Create Image



Code



Summarize



Get advice

More





ChatGT3



Explore GT3s

Today

Project objectives



Research question



Main task



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Message WBS



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

Now, let's move to development. What is the main task a how to set a baseline?



The main task of this phase involves several key decisions as follows:

- **Choosing Datasets:** Select appropriate datasets for training and evaluating the retrieval system.
- **Analyzing Useful Sentence Embedders:** Identify models that effectively convert text into vector representations.
- **Analyzing Useful VectorDBs:** Explore databases optimized for storing and querying vector embeddings.
- **Selecting Similarity Measures:** Choose methods to compare embeddings for accurate retrieval.
- **Selecting Performance Metrics:** Determine metrics to evaluate and benchmark system performance.
- **Forming A Baseline:** Develop an initial system configuration to measure performance and guide improvements.
- **Making Improvements on Top of Baseline:** Iteratively refine the system by testing enhancements to improve the retrieval quality and achieve the defined objectives.

A well-defined baseline is crucial as it provides a reference point to measure progress and evaluate improvements. It ensures that enhancements are meaningful by establishing a solid foundation for comparison and identifying areas of strength or weakness.

Would you like me to also explain the baseline in the same manner as that of the main task?






Message ChatGT3












ADSP - P9 - RAG MARCO.pdf
PDF

Today

- Project objectives 
- Research question 
- Main task 

November

- Project definition 
- Project objectives 
- Project relevance to SDGs 
- Design: user personas 
- Design: customer journey 
- Design: functional diagram 
- Manage WBS 

Now, let's move to development. What is the main task a how to set a baseline?



The main task of this phase involves several key decisions as follows:

- **Choosing Datasets:** Select appropriate datasets for training and evaluating the retrieval system.
- **Analyzing Useful Sentence Embedders:** Identify models that effectively convert text into vector representations.
- **Analyzing Useful VectorDBs:** Explore databases optimized for storing and querying vector embeddings.
- **Selecting Similarity Measures:** Choose methods to compare embeddings for accurate retrieval.
- **Selecting Performance Metrics:** Determine metrics to evaluate and benchmark system performance.
- **Forming A Baseline:** Develop an initial system configuration to measure performance and guide improvements.
- **Making Improvements on Top of Baseline:** Iteratively refine the system by testing enhancements to improve the retrieval quality and achieve the defined objectives.

A well-defined baseline is crucial as it provides a reference point to measure progress and evaluate improvements. It ensures that enhancements are meaningful by establishing a solid foundation for comparison and identifying areas of strength or weakness.

Would you like me to also explain the baseline in the same manner as that of the main task?



Yes, please also explain the baseline. |





Today

Project objectives ...

Research question ...

Main task ...

November

Project definition ...

Project objectives ...

Project relevance to SDGs ...

Design: user personas ...

Design: customer journey ...

Design: functional diagram ...

Message WBS ...



ADSP - P9 - RAG MARCO.pdf
PDF

Yes, please also explain the baseline.



The baseline for the system involves the following steps:

- **Document Chunking:** Break documents into smaller chunks to improve retrieval granularity.
- **Embedding Chunks and Storing in VectorDB:** Convert chunks into vector representations and store them in a vector database for efficient search and retrieval.
- **Taking A Random Sample from Queries and Obtaining Their Embeddings:** Start with a manageable set of approximately 1,000 queries to enable fast and cost-effective trials.
- **Top-K Similar Passage Retrieval for Each Query:** Dynamically retrieve the top-k most relevant passages for each query based on similarity measures.
- **Performance Evaluation:** Assess the retrieval system's effectiveness using predefined performance metrics.



Chunking



Embedding



Queries



Retrieval



Evaluation



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



Research question



Main task



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Design: functional diagram



Manage WBS



Upgrade plan

What can I help with?



ADSP - P9 - RAG MARCO.pdf
PDF



Which datasets are we using, and how do they help power our project? |



Create Image



Code



Summarize



Get advice

More



Today

- Project objectives ...
- Research question ...
- Main task ...
- Datasets ...

November

- Project definition ...
- Project objectives ...
- Project relevance to SDGs ...
- Design: user personas ...
- Design: customer journey ...
- Design: functional diagram ...

Which datasets are we using, and how do they help power our project?



Certainly! Here are the datasets we’re utilizing and how they contribute to the success of our project:

MS-MARCO and **TriviaQA**.

Below, you'll find a comparison between these two datasets.

Datasets	Question Format	Relevant Passages	Selected Passages	Passage Ranking	Generated Answer
MS-MARCO	Query strings asked by users on Bing.	Approximately six passages related to the query.	Approximately two passages selected after retrieval,	No explicit ranking.	A well-formed human-generated answer.
TriviaQA	Trivia questions collected from quiz-league websites.	Approximately six passages containing the answer.	Same as the retrieved passages.	Ranked based on similarity to the question.	Human-verified and machine-generated subsets



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives ...

Research question ...

Main task ...

Datasets ...

November

Project definition ...

Project objectives ...

Project relevance to SDGs ...

Design: user personas ...

Design: customer journey ...

Design: functional diagram ...



Upgrade plan

What can I help with?



ADSP - P9 - RAG MARCO.pdf
PDF



What are embedders and VectorDBs, and how do they aim to improve retrieval? |



Create Image



Code



Summarize



Get advice

More



Today

- Project objectives ...
- Research question ...
- Main task ...
- Datasets ...
- Embedders and VectorDBs ...

November

- Project definition ...
- Project objectives ...
- Project relevance to SDGs ...
- Design: user personas ...
- Design: customer journey ...

What are embedders and VectorDBs, and how do they aim to improve retrieval?





Of course! Let's explore the models we're using to assess similarity and performance:


all-mpnet-base-v2, **multi-qa-mpnet-base-dot-v1**, and **all-distilroberta-v1**.


Below, you'll find an in-depth comparison of their architecture, performance, and areas of application.

Embedders	Architecture	Use Case	Performance
all-mpnet-base-v2	MPNet	General-purpose.	Slow with excellent embedding power
multi-qa-mpnet-base-dot-v1	MPNet	Optimized for semantic search.	Moderate with very good embedding power
all-distilroberta-v1	DistilRoBERTa	General-purpose.	Fast with good embedding power

 Message ChatGT3



 ChatGT3

 Explore GT3s

 ADSP - P9 - RAG MARCO.pdf
PDF

Today

- Project objectives ...
- Research question ...
- Main task ...
- Datasets ...
- Embedders and VectorDBs ...

November

- Project definition ...
- Project objectives ...
- Project relevance to SDGs ...
- Design: user personas ...
- Design: customer journey ...

What are embedders and VectorDBs, and how do they aim to improve retrieval?





Let's take a look at the embedders and VectorDBs we're utilizing to optimize retrieval:

FAISS and **ScaNN**.

Below, you'll find a comprehensive comparison of their features and how they enhance the overall retrieval process.

VectorDB	Algorithm	Customizability	Optimization
FAISS (Facebook AI Similarity Search)	IVF, PQ, HNSW, and Flat indexing.	Tunable balance between accuracy and speed.	GPU acceleration and memory efficiency
ScaNN (Scalable Nearest Neighbors)	Tree-based, ADQ, and Re-ranking.	Focuses on maintaining high accuracy with speed.	CPU-focused optimizations with scalability

 Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



Research question



Main task



Datasets



Embedders and VectorDBs



November

Project definition



Project objectives



Project relevance to SDGs



Design: user personas



Design: customer journey



Upgrade plan

What can I help with?



ADSP - P9 - RAG MARCO.pdf
PDF



How do we measure similarity and on what basis do we want to evaluate it? |



Create Image



Code



Summarize



Get advice

More





ChatGT3



Explore GT3s

Today

Project objectives



Research question



Main task



Datasets



Embedders and VectorDBs



Similarity and performance



November

Project definition



Project objectives



Project relevance to SDGs



Design user persona



Upgrade plan

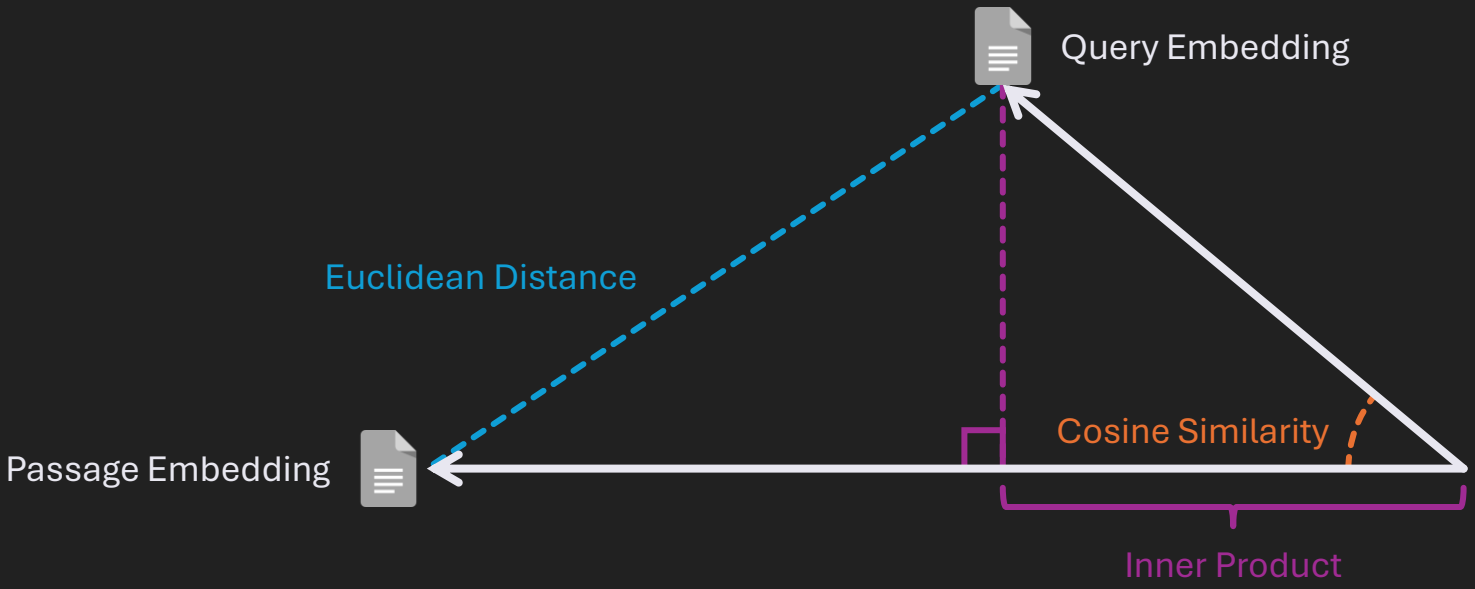


ADSP - P9 - RAG MARCO.pdf
PDF

How do we measure similarity, and why is performance so crucial for this system?



Of course! To measure similarity, we rely on metrics like **L2 distance, Inner product, and Cosine Similarity**, which determine the closeness between vector representations. You can see these metrics illustrated in the following image.



Message ChatGT3





Today

- Project objectives
- Research question
- Main task
- Datasets
- Embedders and VectorDBs
- Similarity and performance

November

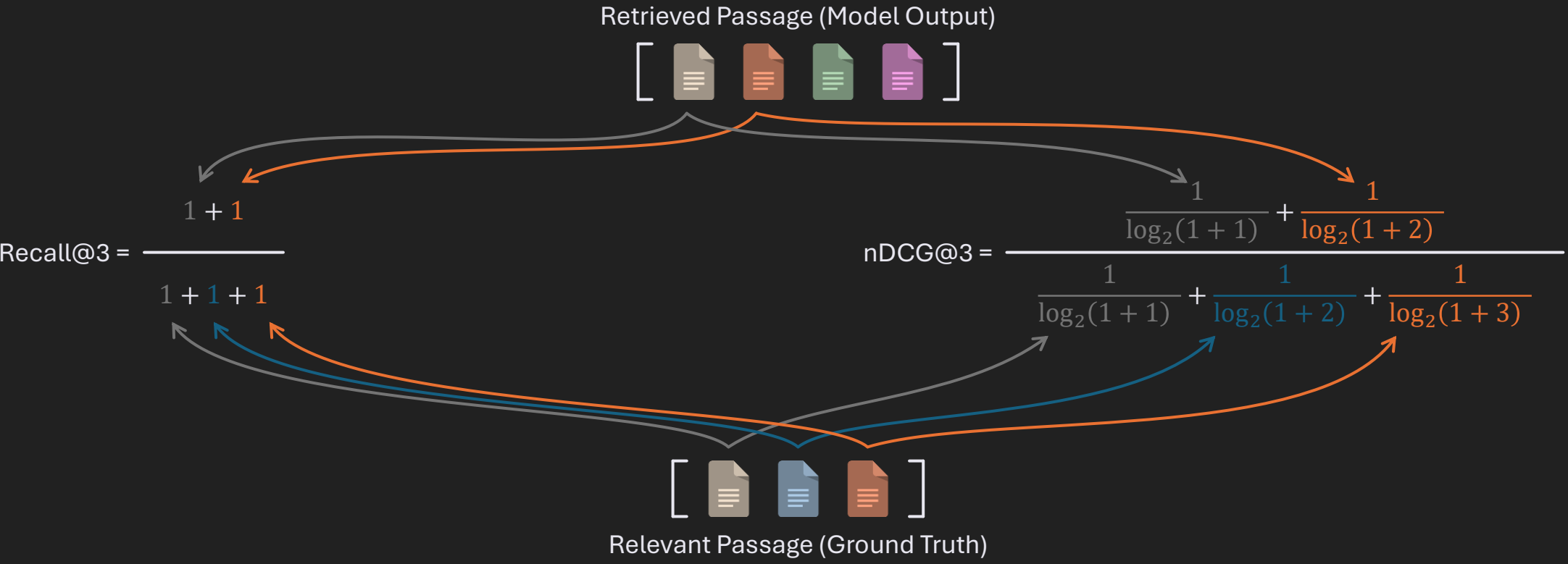
- Project definition
- Project objectives
- Project relevance to SDGs
- Design user persona



How do we measure similarity, and why is performance so crucial for this system?



For evaluating system performance, we use metrics such as **Recall@k** and **nDCG@k**, ensuring relevance and ranking quality in retrieved results. Below, you'll find a comparison of these similarity measures and performance metrics.



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives ...

Research question ...

Main task ...

Datasets ...

Embedders and VectorDBs ...

Similarity and performance ...

November

Project definition ...

Project objectives ...

Project relevance to SDGs ...

Design user personas ...



Upgrade plan

What can I help with?



ADSP - P9 - RAG MARCO.pdf
PDF



How will we continue pushing this project forward? |



Create Image



Code



Summarize



Get advice

More





Today

Project objectives ...

Research question ...

Main task ...

Datasets ...

Embedders and VectorDBs ...

Similarity and performance ...

Next steps ...

November

Project definition ...

Project objectives ...

Project relevance to SDOs ...



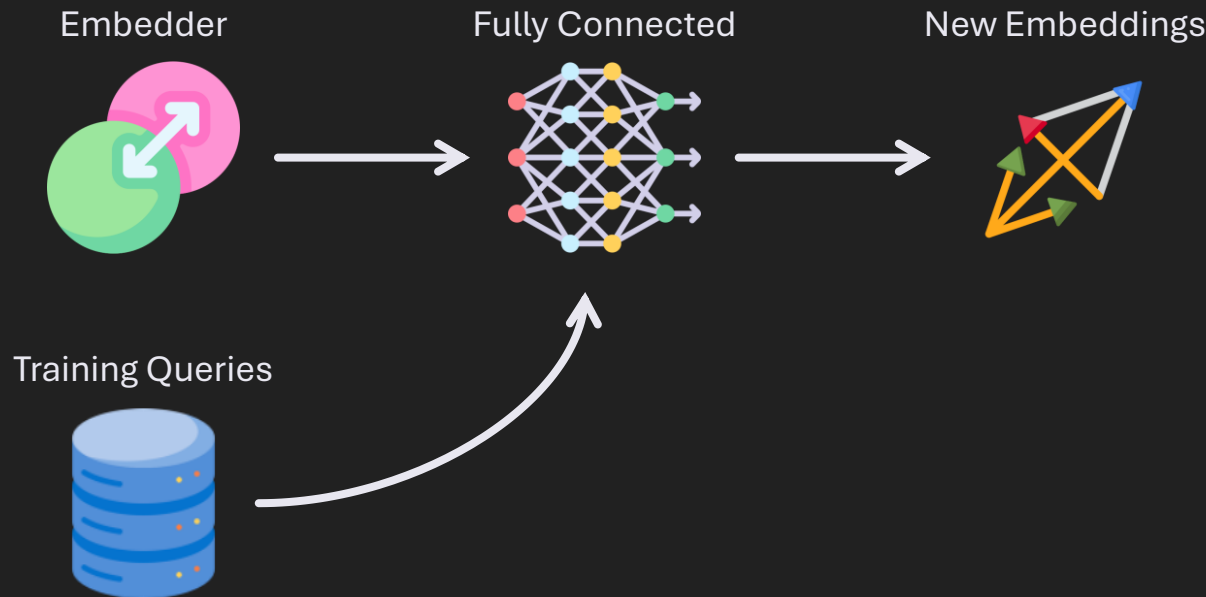
ADSP - P9 - RAG MARCO.pdf
PDF

How will we continue pushing this project forward?



Applying a **fully connected layer** to the current embedder models and training it using the available datasets.

- **Enriching the embeddings** to bring query and document embeddings **closer together in the embedding space**, which leads to improved retrieval accuracy.
- **Reducing the dimensionality of the representations** below the original value to provide **compression capability**, which leads to lower computational costs.



Message ChatGT3





Today

- Project objectives
- Research question
- Main task
- Datasets
- Embedders and VectorDBs
- Similarity and performance
- Next steps

November

- Project definition
- Project objectives
- Project relevance to SDOs



How will we continue pushing this project forward?



On top of the **semantic similarity scores** measured by the fully connected architecture, we can also employ NER, Keyword, and Topic extraction to obtain **syntactic similarity scores** through applying the following measures:

- NER Extraction:** Identify the named entities such as persons, organizations, locations, or dates, which can be achieved using spaCy or NLTK libraries.
- Keyword Extraction:** Extract specific keywords using YAKE library.
- Topic Extraction:** Determine the main theme of passages and queries.

$$S = w_1 S_{FC} + w_2 S_{NER} + w_3 S_{KEYWORD} + w_4 S_{TOPIC}$$

Where, **for the weights, we have:**

$$w_1 + w_2 + w_3 + w_4 = 1$$

Also, **for the components scores, we have:**

$$S_{NER} = \frac{|\mathbb{E}_Q \cap \mathbb{E}_P|}{|\mathbb{E}_Q|}, \quad S_{KEYWORD} = \frac{|\mathbb{K}_Q \cap \mathbb{K}_P|}{|\mathbb{K}_Q|}, \quad S_{TOPIC} = \begin{cases} 0 & t_Q \neq t_P \\ 1 & t_Q = t_P \end{cases}$$



Message ChatGT3





How will we continue pushing this project forward?



Moreover, **instead of manually tuning the weights**, w_1 to w_4 , we can utilize a **transformer architecture to learn them** dynamically as attention scores. The process of this attention block is as follows:

- **Keys (K)**: The retrieved passages.
- **Queries (Q)**: The input queries.
- **Values (V)**: The similarity scores computed for each passage (S_{FC} , S_{NER} , $S_{KEYWORD}$, and S_{TOPIC}).

Note that, in the attention mechanism, the transformer computes attention scores that represent the relative importance of the similarity components, which correspond to w_1 to w_4 according the following formula:

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \text{softmax} \left(\frac{Q \times K^T}{\sqrt{d_k}} \right) V$$



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



Research question



Main task



Datasets



Embedders and VectorDBs



Similarity and performance



Next steps



November

Project definition



Project objectives



Project relevance to SDOs



Upgrade plan

What can I help with?



ADSP - P9 - RAG MARCO.pdf
PDF



What have we learned, and how will this impact the future of RAG systems? |



Create Image



Code



Summarize



Get advice

More





Today

Project objectives ...

Research question ...

Main task ...

Datasets ...

Embedders and VectorDBs ...

Similarity and performance ...

Next steps ...

Conclusion ...

November

Project definition ...

Project objectives ...



ADSP - P9 - RAG MARCO.pdf
PDF

What have we learned, and how will this impact the future of RAG systems?



In conclusion, to answer this question, our findings highlight the importance of balancing semantic richness and computational efficiency to achieve state-of-the-art performance in retrieval-augmented generation (RAG) systems. This research emphasizes that a holistic approach, combining advances in neural network architectures and fine-tuned retrieval strategies, is critical for the next generation of intelligent systems.

- Our work is focused on **enhancing the quality of retrieval in RAG systems** by leveraging a combination of techniques to better **align query embeddings with relevant passages**.
- We began by introducing the **main task** and established a **solid baseline**, utilizing **vector databases, similarity measures**, and **key performance metrics** to evaluate retrieval quality.
- We have proposed a number of strategies to take **advantage of both syntactic and semantic similarities** while also utilizing **attention mechanism**.
- Through the integration of deep learning, natural language processing, and traditional information retrieval techniques, **we are hopeful** that they will lead to improvements in the performance metrics utilized in this project.



Message ChatGT3





ChatGT3



Explore GT3s

Today

Project objectives



Research question



Main task



Datasets



Embedders and VectorDBs



Similarity and performance



Next steps



Conclusion



November

Project definition



Project objectives



Upgrade plan



ADSP - P9 - RAG MARCO.pdf
PDF

What have we learned, and how will this impact the future of RAG systems?



In conclusion, to answer this question, our findings highlight the importance of balancing semantic richness and computational efficiency to achieve state-of-the-art performance in retrieval-augmented generation (RAG) systems. This research emphasizes that a holistic approach, combining advances in neural network architectures and fine-tuned retrieval strategies, is critical for the next generation of intelligent systems.

- Our work is focused on **enhancing the quality of retrieval in RAG systems** by leveraging a combination of techniques to better **align query embeddings with relevant passages**.
- We began by introducing the **main task** and established a **solid baseline**, utilizing **vector databases, similarity measures**, and **key performance metrics** to evaluate retrieval quality.
- We have proposed a number of strategies to take **advantage of both syntactic and semantic similarities** while also utilizing **attention mechanism**.
- Through the integration of deep learning, natural language processing, and traditional information retrieval techniques, we have developed a comprehensive framework that is **tailored and optimized** for this project.

You've hit the Free plan limit for GT3.
You need GT3-4o to continue this chat because there's an attachment.
Your limit resets after January 9th.

Get Plus



Message ChatGT3





Oops!

Thank you for your attention, but our systems are busy at the moment.
Please take a break and ask your questions.



Homayoun Afshari



Arash Daneshvar



Hossein Khodadadi

You've hit the Free plan limit for GT3.

You need GT3-4o to continue this chat because there's an attachment.
Your limit resets after January 9th.

Get Plus