

Master thesis  
Toward Optimizing  
a Retrieval  
Augmented  
Generation Pipeline  
using Large  
Language Model

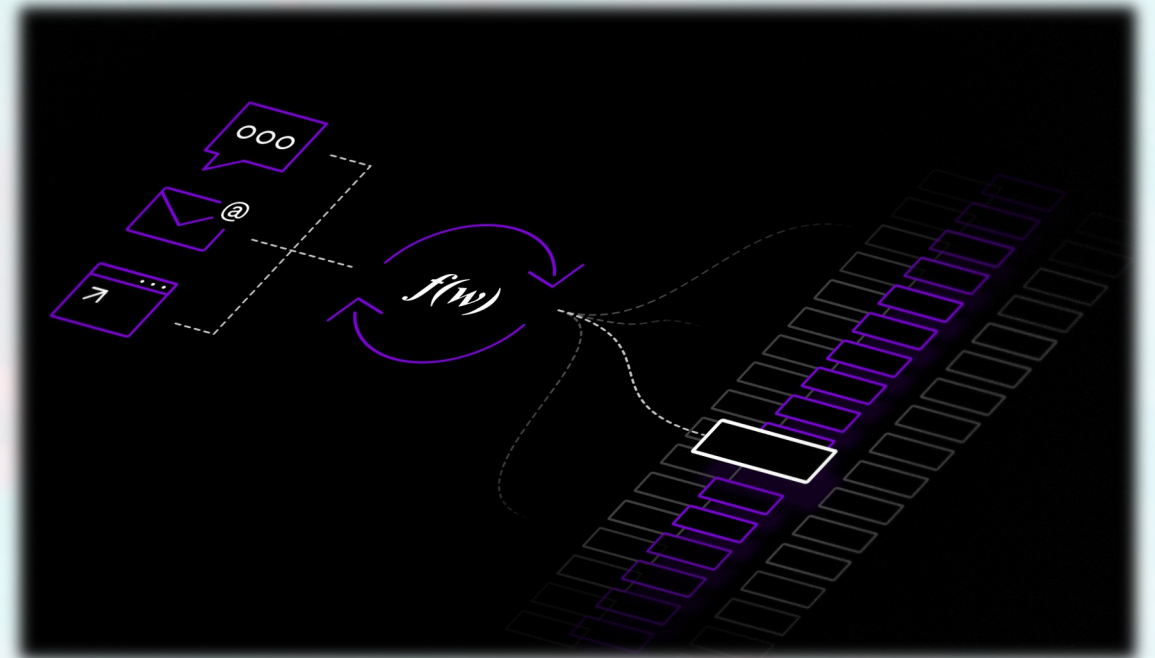
Student: Gentrit Fazlija

Supervisor: Anum Afzal

Professor: Prof. Dr. Florian Matthes

# Table of content

- i. Motivation
- ii. Corpora
- iii. Modular RAG
- iv. Experiments
- v. Evaluation
- vi. Conclusion & Future Work



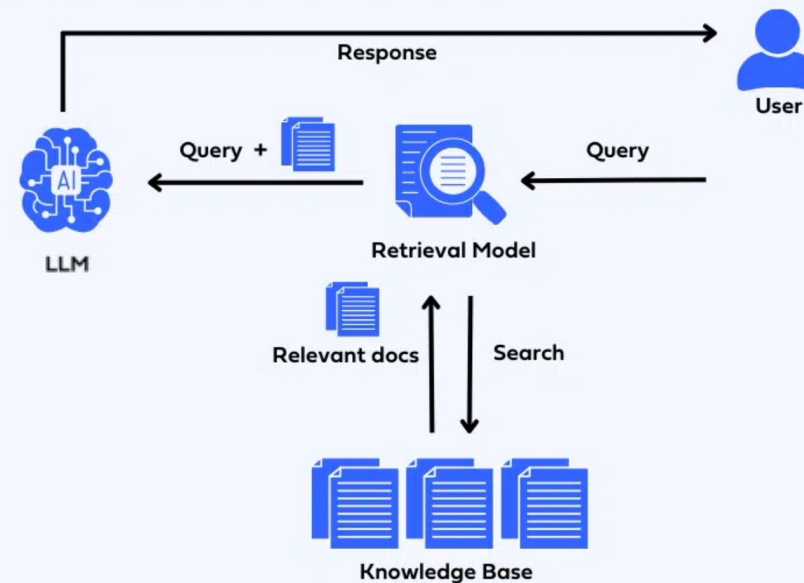
# Motivation

RAG framework allows:

- Connect real time data
- Update current corpora
- Enable or disable knowledge flow
- Interpret Hit Rate
- Possibly less computation

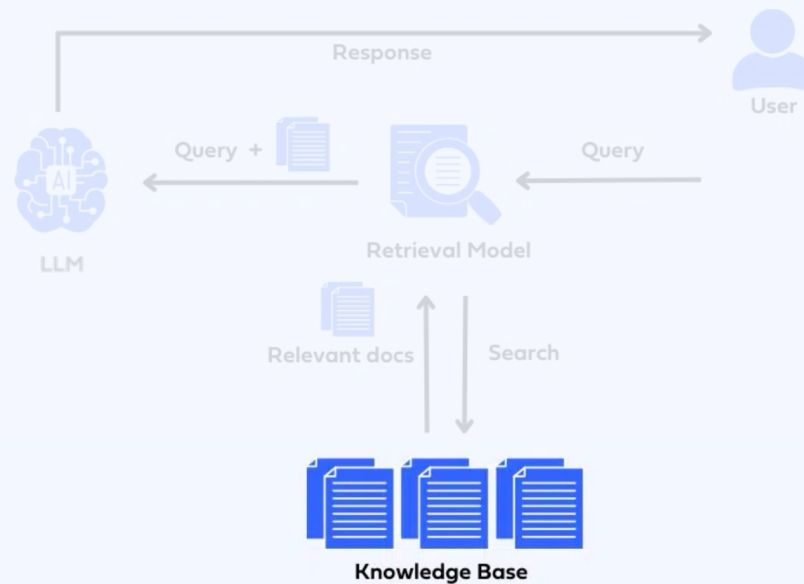
→ Databricks study: approximately 60%

## Retrieval Augmented Generation



# Motivation

## Retrieval Augmented Generation



Digestion

# Knowledge Base

Clear and easy scraping

All studyprograms have the same structure

Official TUM website

Master of Science (M.Sc.)  
**Mathematics in Data Science**

The master's degree program Mathematics in Data Science combines a high-profile education in mathematics with an emphasis on the burgeoning area of Big Data.

[TUM School of Computation, Information and Technology](#)

**Key Data**

<b>Type of Study</b> Full Time	<b>Standard Duration of Studies</b> 4 (fulltime)	<b>Credits</b> 120 ECTS
<b>Main Locations</b> <a href="#">Garching</a>	<b>Application Period</b> Winter semester: 01.01. – 31.05. Summer semester: 01.09. – 30.11.	<b>Admission Category</b> <a href="#">Aptitude Assessment for Master</a>
<b>Start of Degree Program</b> Possible for both winter and summer semester	<b>Costs</b> <a href="#">Student Fees: 85.00 €</a>	<b>Required Language Proficiency</b> English

**Information on Degree Program**

- What does this program cover? +
- How is the program structured? +
- What is the language of instruction? +
- Which further expertise and skills will I acquire? +

**Binding Regulations for Progression of Studies, Examinations and Application**

<https://www.tum.de/en/studies/degree-programs>

# Knowledge Base

Individual TUM faculty  
website

Here: Clear headings

All studyprograms have  
different structures

## Master Mathematics in Data Science

[Beratung](#) ▾

[Das Wichtigste zum Masterstudium](#) ▾

[Voraussetzungen für den Master Mathematik](#) ▾

[Bewerbung für den Master](#) ▾

[Das Wichtigste zur Masterarbeit](#) ▾

Sie jonglieren gerne mit Daten wie sie etwa in den sozialen Medien generiert werden? Sie finden Verfahren zur Datenerfassung, Datenaufbereitung und Datenanalyse richtig spannend und wollen diese bei komplexen Daten in ökologischen Systemen anwenden? Das Masterstudium "Mathematics in Data Science" an der Technischen Universität München (TUM) macht Sie fit für den zukunftsträchtigen Arbeitsbereich Big Data. Als Daten-Experte sind Sie eine gefragte Person in Forschung und Entwicklung, in der Finanzindustrie, Biotechnologie und Logistik, im Gesundheitswesen, bei Versicherungen und für IT-Sicherheit.

Fachstudienberater



**PD Peter Massopust,**  
**Ph.D.**  
[mscapp\\_datascience](mailto:mscapp_datascience@ma.tum.de)  
[\(at\) ma.tum.de](mailto:m(at)ma.tum.de)

### Studienschwerpunkte

Im Masterprogramm "Mathematics in Data Science" stehen die Bereiche Data Engineering, Data Analytics, Data Analysis, Machine Learning und Data Science im Mittelpunkt. Studierende konzentrieren sich dabei auf Techniken der Datenhaltung und -auswertung. Sie lernen, diese an konkrete Problemstellungen anzupassen, sie zu kombinieren oder neu zu entwickeln und daraus Vorhersage- und Klassifikationsmodelle abzuleiten. Daneben beschäftigen sie sich auch mit der Weiterentwicklung von Algorithmen zur Problemlösung. Überfachliche Lehrveranstaltungen, die sich mit gesellschaftlichen und politischen Implikationen von Big Data beschäftigen sowie juristisches Grundwissen und Fremdsprachenkenntnisse vermitteln, sind ebenso wichtige Bestandteile des Masterprogramms.

### Nicht verpassen!

Hier finden Sie wichtige Termine, Deadlines und andere relevante Hinweise.



#### Bewerbungsfristen

Wer mit dem Masterstudium beginnen will, sollte sich möglichst frühzeitig bewerben. Alle notwendigen Unterlagen müssen Sie innerhalb der Fristen in Ihrem Bewerberaccount im [Online-System der TUM](#) vollständig hochladen.

**Bewerbungsfrist für das Wintersemester: 1. Januar bis 31. Mai**

**Bewerbungsfrist für das Sommersemester: 1. September bis 30. November**

# Knowledge Base

Here: Not clear headings

All studyprograms have different structures

Individual TUM faculty website

The screenshot shows a website page for the Master in Finance & Information Management program at TUM. The header includes the TUM School of Management logo and navigation links: PROGRAMS, CAMPUSES, FACULTY & RESEARCH, CORPORATE PARTNERS, ABOUT, and OUR STORIES. A breadcrumb trail reads: mgittum.de > Programs > Graduate programs > Master in Finance & Information Management. The main content area features a large image of a woman presenting in a classroom. Overlaid on the image is the text: "MASTER IN FINANCE & INFORMATION MANAGEMENT | MUNICH" and "NAVIGATE FINANCE AND DIGITALIZATION IN A DATA-DRIVEN WORLD". A white box on the right contains the program details:

## Master in Finance & Information Management

- Location: Munich
- Language: 100% in English
- Full-time study
- 4 semesters
- Certificate of honors
- No tuition fees until summer semester 2025. Detailed information can be found [here](#).
- Start of program: Winter semester
- Admission period:
  - 1st application period from **January 1 – March 15**

# Knowledge Base

## Problem Identification:

- Semantically similar in wording and content
- Difficulty to distinguish the information
- Context-awareness is mainly focused in the headings

Master of Science (M.Sc.)

## Mathematics in Data Science

The master's degree program Mathematics in Data Science combines a high-profile education in mathematics with an emphasis on the burgeoning area of Big Data.

TUM School of Computation, Information and Technology [🔗](#)

Key Data		
Type of Study Full Time	Standard Duration of Studies 4 (fulltime)	Credits 120 ECTS
Main Locations Garching	Application Period Winter semester: 01.01. – 31.05. Summer semester: 01.09. – 30.11.	Admission Category Aptitude Assessment for Master
Start of Degree Program Possible for both winter and summer semester	Costs Student Fees: 85.00 €	Required Language Proficiency English

Information on Degree Program

- What does this program cover? +
- How is the program structured? +
- What is the language of instruction? +

## Master Mathematics in Data Science

Beratung [▼](#)


Das Wichtigste zum Masterstudium [▼](#)

Voraussetzungen für den Master Mathematik [▼](#)

Bewerbung für den Master [▼](#)

Das Wichtigste zur Masterarbeit [▼](#)

Sie jonglieren gerne mit Daten wie sie etwa in den sozialen Medien generiert werden? Sie finden Verfahren zur Datenerfassung, Datenaufbereitung und Datenanalyse richtig spannend und wollen diese bei komplexen Daten in biologischen Systemen anwenden? Das Masterstudium "Mathematics in Data Science" an der Technischen Universität München (TUM) macht Sie fit für den zukunftsreichen Arbeitsbereich Big Data. Als Daten-Experte sind Sie eine gefragte Person in Forschung und Entwicklung, in der Finanzindustrie, Biotechnologie und Logistik, im Gesundheitswesen, bei Versicherungen und für IT-Sicherheit.

Fachstudienberater  

 PD Peter Massopust, Ph.D.  
[massopust@ma.tum.de](mailto:massopust@ma.tum.de)

### Studienschwerpunkte

Im Masterprogramm "Mathematics in Data Science" stehen die Bereiche Data Engineering, Data Analytics, Data Analysis, Machine Learning und Data Science im Mittelpunkt. Studierende konzentrieren sich dabei auf Techniken der Datenhaltung und -auswertung. Sie lernen, diese an konkrete Problemstellungen anzupassen, sie zu kombinieren oder neu zu entwickeln und daraus Vorhersage- und Klassifikationsmodelle abzuleiten. Daneben beschäftigen sie sich auch mit der Weiterentwicklung von Algorithmen zur Problemlösung. Überfachliche Lehrveranstaltungen, die sich mit gesellschaftlichen und politischen Implikationen von Big Data beschäftigen sowie juristisches Grundwissen und Fremdsprachenkenntnisse vermitteln, sind ebenso wichtige Bestandteile des Masterprogramms.

### Nicht verpassen!

Hier finden Sie wichtige Termine, Deadlines und andere relevante Hinweise.

**Bewerbungsfristen**

Wer mit dem Masterstudium beginnen will, sollte sich möglichst frühzeitig bewerben. Alle notwendigen Unterlagen müssen Sie innerhalb der Fristen in Ihrem Bewerberaccount im [Online-System der TUM](#) vollständig hochladen.

Bewerbungsfrist für das Wintersemester: 1. Januar bis 31. Mai  
 Bewerbungsfrist für das Sommersemester: 1. September bis 30. November

TUM School of Management  
Technical University of Munich

PROGRAMS | CAMPUSES | FACULTY & RESEARCH | CORPORATE PARTNERS | ABOUT | OUR STORIES

ingtudie > Programs > Graduate programs > Master in Finance & Information Management

## Master in Finance & Information Management

NAVIGATE FINANCE AND DIGITALIZATION IN A DATA-DRIVEN WORLD

- Location: Munich
- Language: 100% in English
- Full-time study
- 4 semesters
- Certificate of honors
- No tuition fees until summer semester 2025. Detailed information can be found [here](#).
- Start of program: Winter semester
- Admission period:
  - 1st application period from January 1 – March 15
  - 2nd application period March 16 – May 31

Interested in this program? [INTERESTED IN THIS PROGRAM](#) | [A CURRENT STUDENT](#)

Overview FM



# Knowledge Base

Root Key = Studyprogram

150+ study programs

6326 total topics

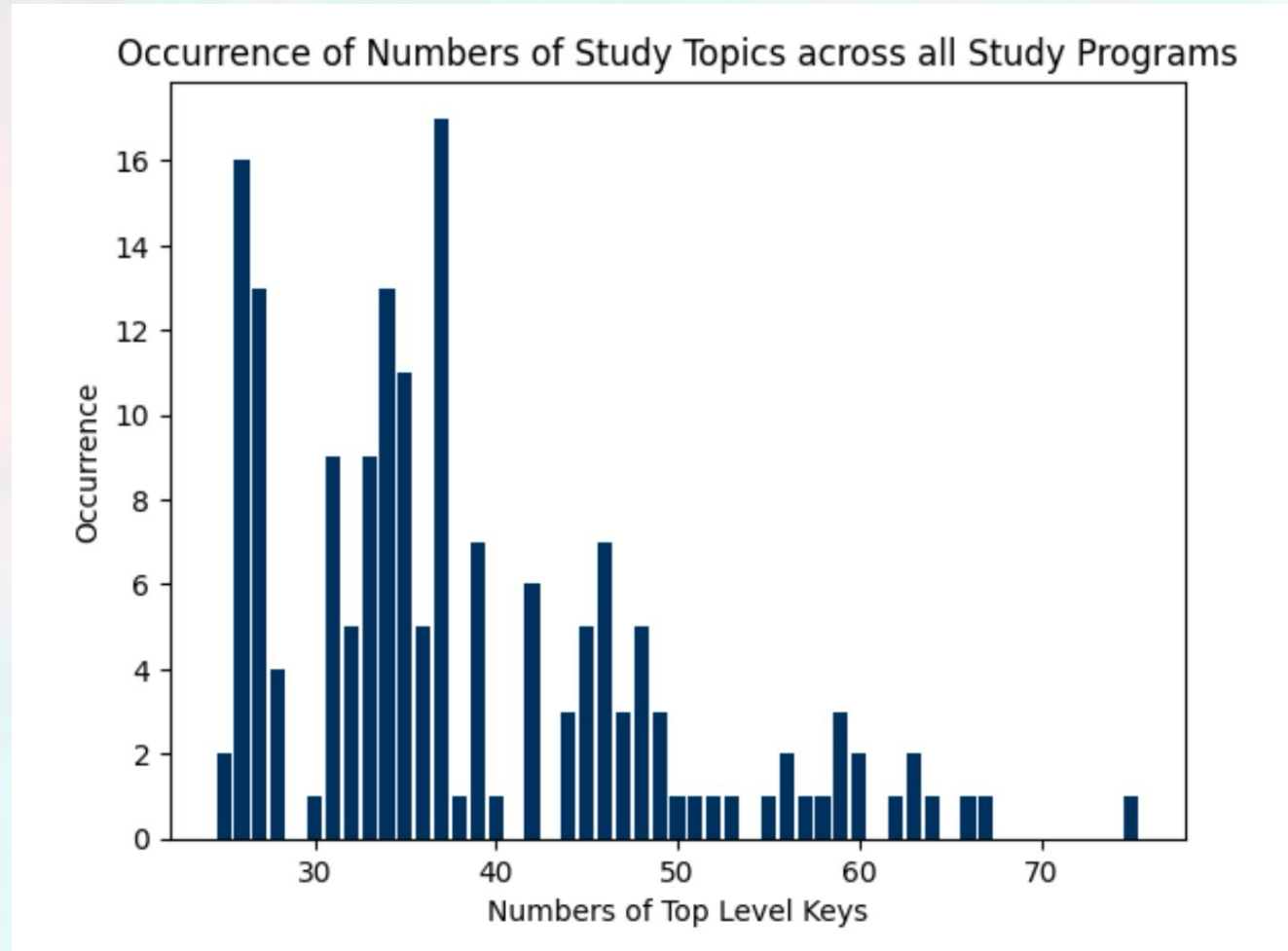
DE & EN

Top Level Key = Topic

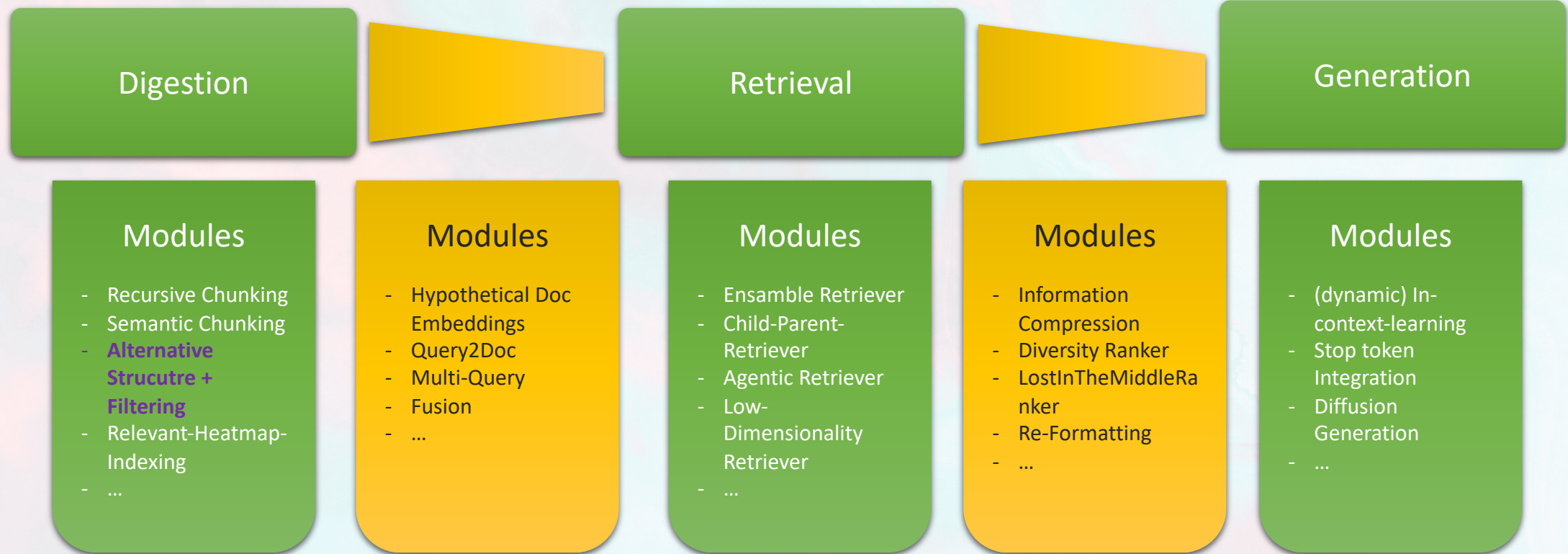
Provided a small sample of value to optimize topic description

```
{
  "Biogenic Materials Bachelor of Science (B.Sc.)": {
    "level": "Bachelor of Science (B.Sc.)",
    "studiengang": "Biogenic Materials",
    "description": "The finiteness of petroleum and fossil-based raw materials and structural changes towards a bioeconomy require the development of innovative biogenic materials. A particular challenge is combining the requirements regarding processability and high-performance properties with sustainability aspects.",
    "school": "The finiteness of petroleum and fossil-based raw materials and structural changes towards a bioeconomy require the development of innovative biogenic materials. A particular challenge is combining the requirements regarding processability and high-performance properties with sustainability aspects.",
    "school_website": "https://www.cs.tum.de/studieninteressierte/studiengaenge/biogene-werkstoffe/",
    "Type of Study": "Full Time",
    "Standard Duration of Studies": "6 semesters (fulltime)",
    "Credits": "180 ECTS",
    "Main Locations": "Straubing",
    "Application Period": "Winter semester: 15.05. – 15.07.",
    "Admission Category": "Unrestricted Admission",
    "Start of Degree Program": "Winter Semester (October)",
    "Costs": "Student Fees: 72.00 €, Tuition fees for international students",
    "Required Language Proficiency": "German",
    "Program profile": "In the light of the dramatic environmental and climate changes in recent years and the finiteness of fossil raw materials like petroleum, material sciences specializing in biogenic materials is seen as a progressive and visionary discipline. The course “Biogenic Materials” provides specific knowledge about biodegradable, eco-friendly, or environmental damage-reducing materials obtained from renewable resources. The course also creates awareness of sustainable use of resources among the future experts in material science. The technical details of extraction, transformation, shaping and processing, possible areas of application and disposal as well as particular structure-property interrelation of materials of all classes are further aspects that are covered in the course
```

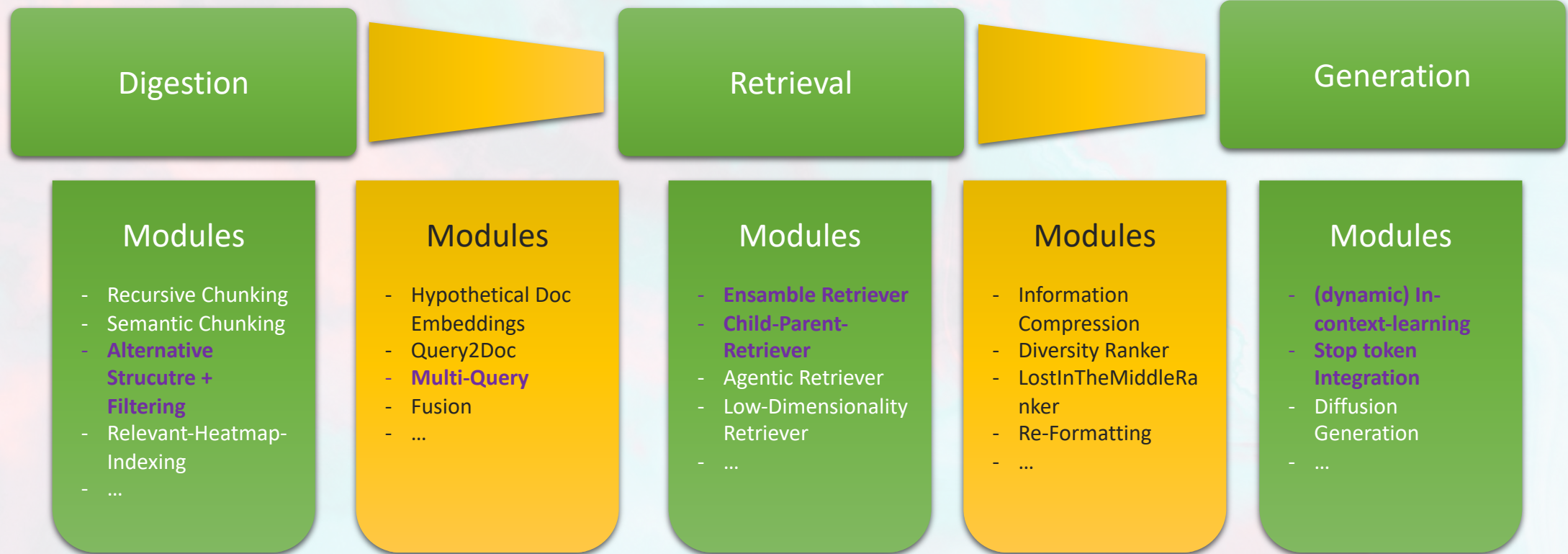
# Knowledge Base



# Modular RAG



# Modular RAG



# Modular RAG on LLMs

Closed

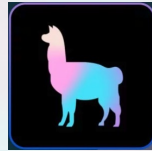


GPT 3.5

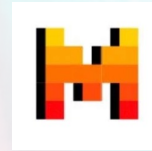


GPT 4

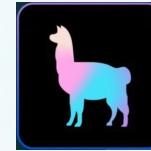
Open



Llama 2 (7B)



Mistral (7B)



Llama 2 (13B)

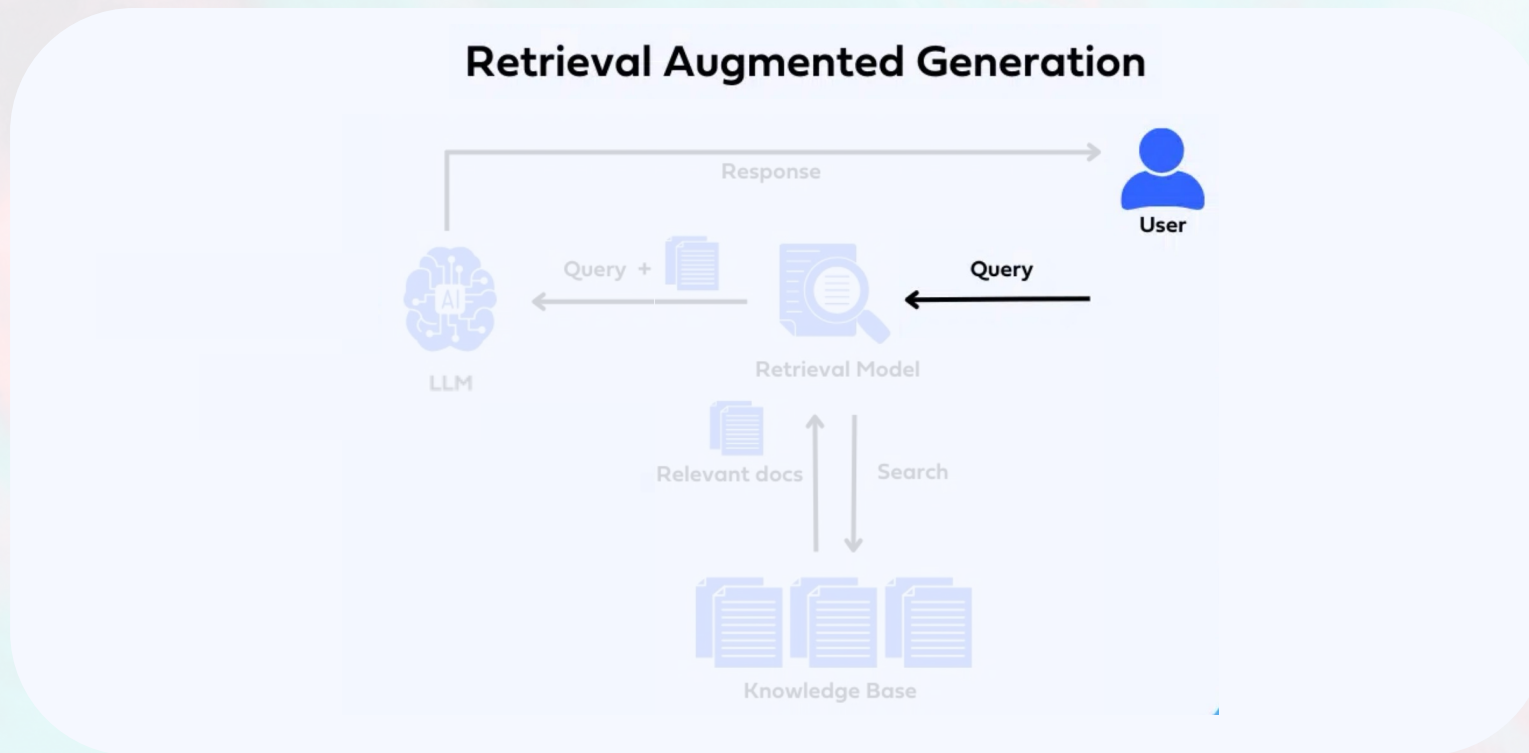


Mini Orca (7B)



Vicuna (7B)

# Key Components & Motivation



Digestion



Retrieval

# Research Question

**RQ 1:** Would a multi-query formulation system improve the performance?

**RQ 2:** Would an optimization approaches, such as ensemble retriever in combination with a child-parent chunking improve the performance of the passage retriever?

**RQ 3:** Would few-shot learning enhance the performance of the system as compared to the Zero-Shot of the system?

**RQ 4:** How does the performance change when using an open-source model compared to a paid closed source model? How can open-sourced models be optimized?

# Challenge & Solution

## Challenge:

Students do not formulate the questions exactly



Students ask the same question in different manners



They might not even know exactly what they are looking for

## Solution:





# Challenge & Solution

## Challenge:

Students do not formulate the questions exactly

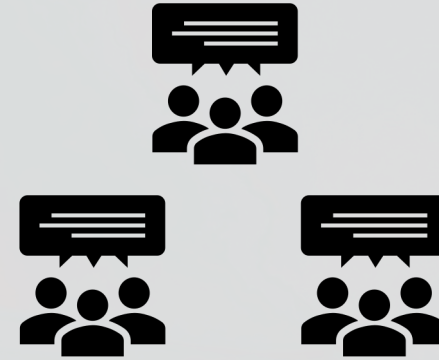


Students ask the same question in different manners



They might not even know exactly what they are looking for

## Solution:



# Challenge & Solution

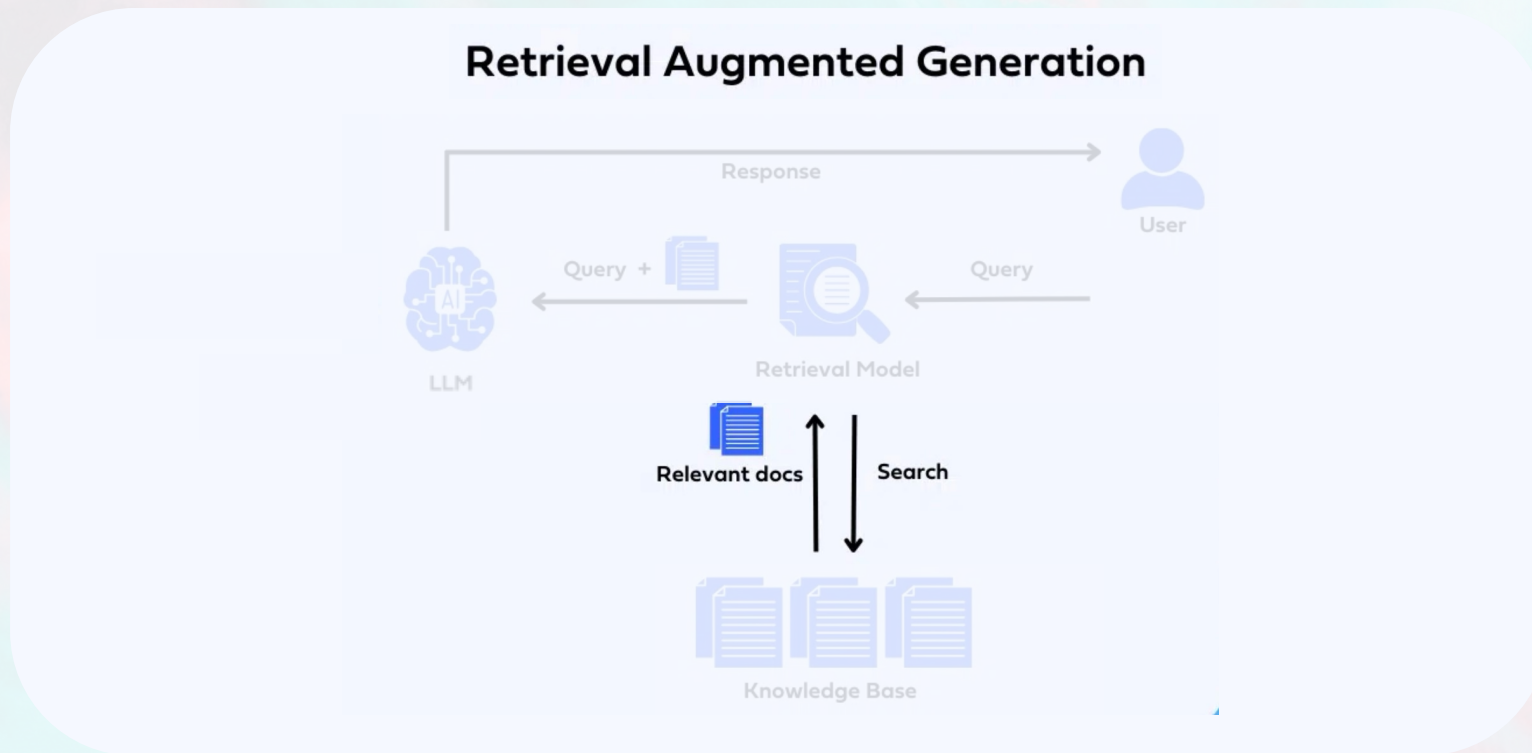
## Solution:



```
Your task is to generate two other different versions of the given user question to retrieve relevant documents from a vector database with respect to the areas of interest. By generating multiple perspectives on the user question, your goal is to help the user overcome some of the limitations of the distance-based similarity search. Provide these alternative questions numbered from 1. to 2. in newlines.
```

```
Context: The Studyprogram is called '{studyprogram}' and the area of interest are '{top_level_keys}'
```

# Key Components & Motivation



Retrieval

# Research Question

**RQ 1:** Would a multi-query formulation system improve the performance?

**RQ 2:** Would an optimization approach, such as ensemble retriever in combination with a child-parent chunking improve the performance of the passage retriever?

**RQ 3:** Would few-shot learning enhance the performance of the system as compared to the Zero-Shot of the system?

**RQ 4:** How does the performance change when using an open-source model compared to a paid closed source model? How can open-sourced models be optimized?

# Challenge & Solution

## Challenge:

Small Chunks: Does not capture the whole meaning



Big Chunks: Has difficulty with similarity search

## Solution:

1

### Child-Chunk

Feld, Feldübungen etc. mit dem Ziel der Durchführung, Auswertung und Erkenntnisgewinnung. <sup>2</sup>Bestandteil können z. B. sein: die Beschreibung der Vorgänge und die jeweiligen theoretischen Grundlagen inkl. Literaturstudium, die Vorbereitung und praktische Durchführung, ggf. notwendige Berechnungen, ihre Dokumentation und Auswertung sowie die Deutung der Ergebnisse hinsichtlich der zu erarbeitenden



### Parent-Chunk

b) <sup>1</sup>Eine **Laborleistung** beinhaltet je nach Fachdisziplin Versuche, Messungen, Arbeiten im Feld, Feldübungen etc. mit dem Ziel der Durchführung, Auswertung und Erkenntnisgewinnung. <sup>2</sup>Bestandteil können z. B. sein: die Beschreibung der Vorgänge und die jeweiligen theoretischen Grundlagen inkl. Literaturstudium, die Vorbereitung und praktische Durchführung, ggf. notwendige Berechnungen, ihre Dokumentation und Auswertung sowie die Deutung der Ergebnisse hinsichtlich der zu erarbeitenden Erkenntnisse. <sup>3</sup>Die Laborleistung kann durch eine Präsentation ergänzt werden, um die kommunikative Kompetenz bei der Darstellung von wissenschaftlichen Themen vor einer Zuhörerschaft zu überprüfen.

# Challenge & Solution

## Solution:

1

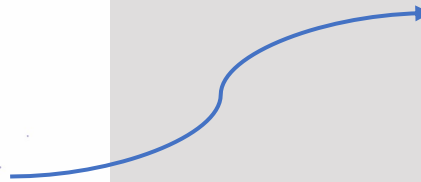
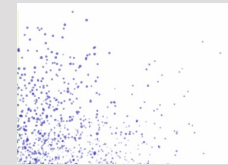
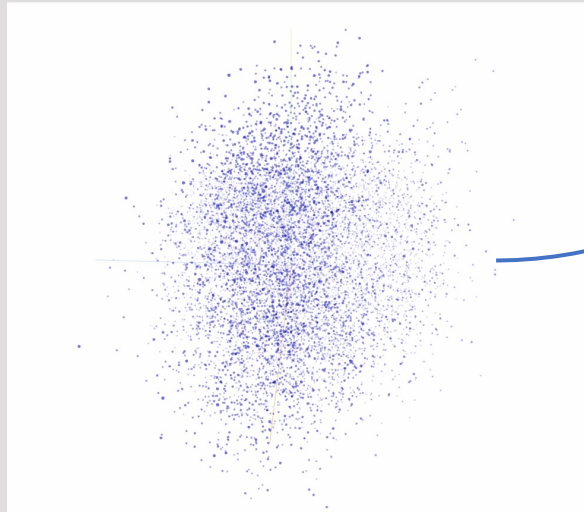
Child-Chunk

Das Foto zeigt die ersten 1000 Bilder der Kamera, die in der ersten Phase der Aufnahme verwendet werden. Die Bilder sind in einem Raster angeordnet und zeigen die ersten 1000 Bilder der Kamera, die in der ersten Phase der Aufnahme verwendet werden. Die Bilder sind in einem Raster angeordnet und zeigen die ersten 1000 Bilder der Kamera, die in der ersten Phase der Aufnahme verwendet werden.



Parent-Chunk

Die Eltern-Chunk sind die ersten 1000 Bilder der Kamera, die in der ersten Phase der Aufnahme verwendet werden. Die Eltern-Chunk sind die ersten 1000 Bilder der Kamera, die in der ersten Phase der Aufnahme verwendet werden. Die Eltern-Chunk sind die ersten 1000 Bilder der Kamera, die in der ersten Phase der Aufnahme verwendet werden.



We connect the parent chunks via ID code with the child chunks

# Challenge & Solution

## Challenge:

Data is very similar in between  
different study programs

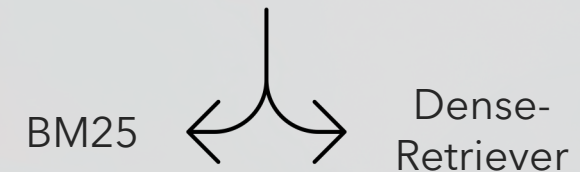
+

Data within one study program  
has specific words being specific  
things

## Solution:

2

Retrieval-System

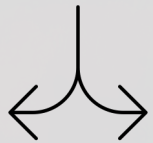


# Challenge & Solution

## Solution:

2

Retrieval-  
System



BM 25    Dense-  
Retriever



```
Q: What methods and models do students in the Bioökonomie Bachelor of Science program learn for evaluating the effects of biobased innovations on companies, the economy, and society?
```

Special Words:

studyprogram

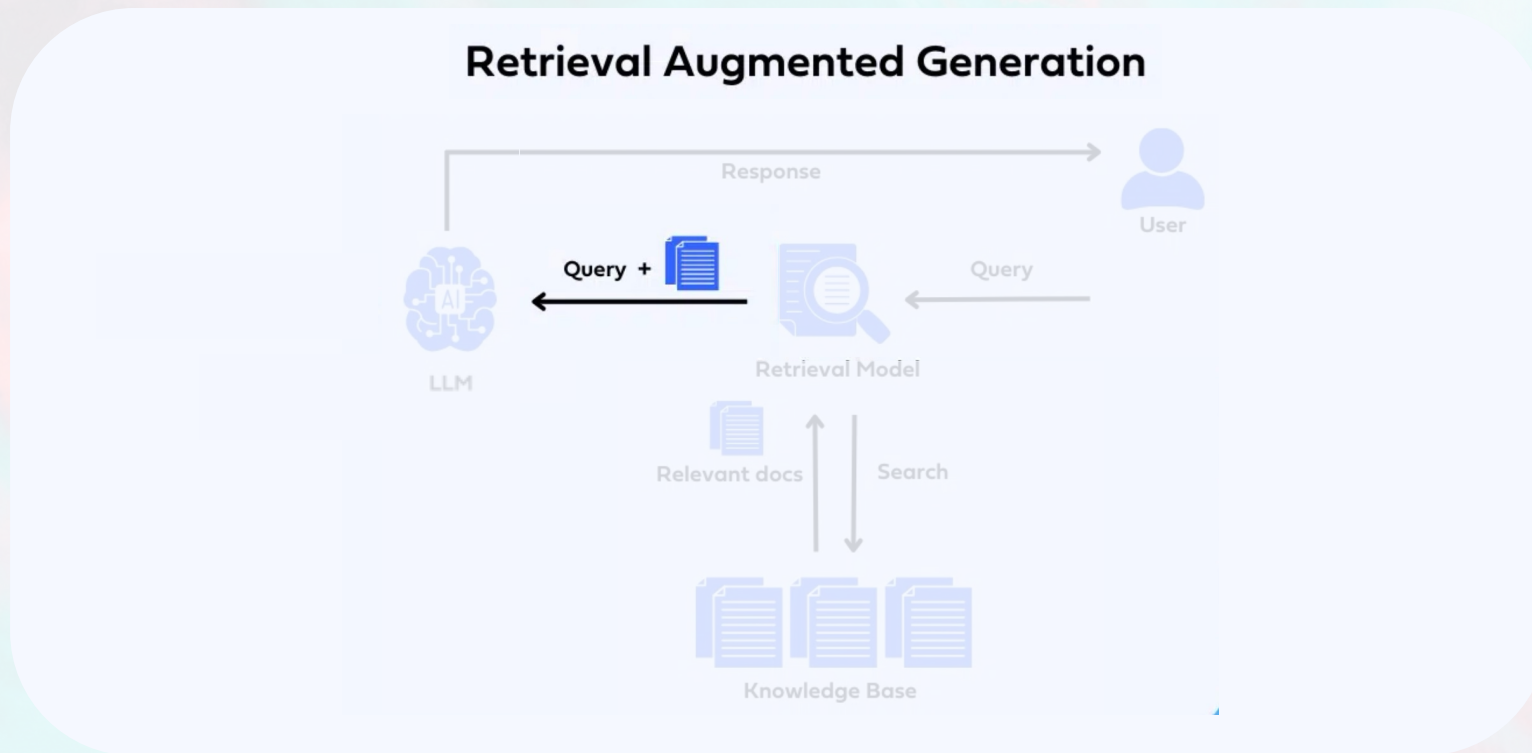
topic area:

- Semantic similarity is not important

- IF the question is clean, then big hit rate

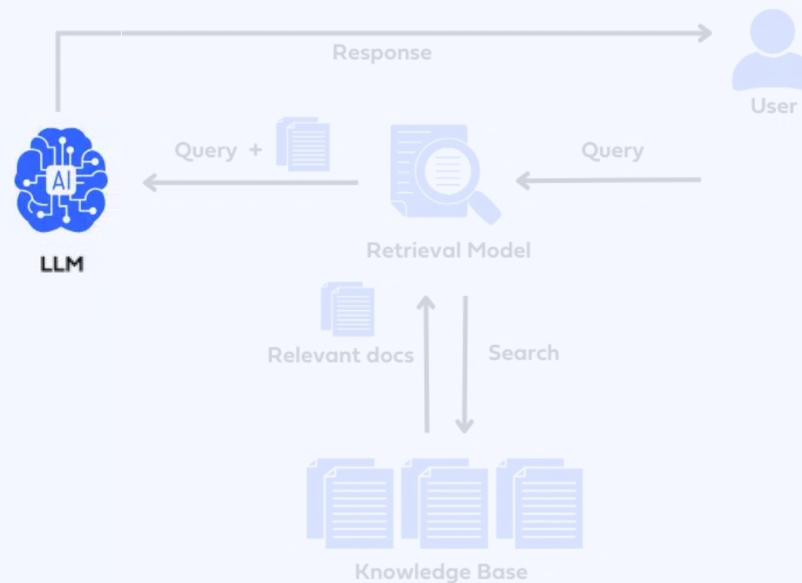


# Key Components & Motivation



# Key Components & Motivation

## Retrieval Augmented Generation



Generation

# Research Question

**RQ 1:** Would a multi-query formulation system improve the performance?

**RQ 2:** Would an optimization approaches, such as ensemble retriever in combination with a child-parent chunking improve the performance of the passage retriever?

**RQ 3:** Would few-shot learning enhance the performance of the system as compared to the Zero-Shot of the system?

**RQ 4:** How does the performance change when using an open-source model compared to a paid closed source model? How can open-sourced models be optimized?

# Challenge & Solution

## Challenge:

LLMs tend to hallucinate

+

LLMs tend to give highly divergent outputs

## Solution:

"What is ..."



ICL

+

"What is ..."

# Challenge & Solution

## Solution:

"What is ..."



ICL



"What is ..."

Answer the question in one to two sentences based only on the following context.

CONTEXT: `{{context}}`

**ICL:**

**Q1:** How do I apply for the Masters program in Management at TUM if I have an undergraduate degree from outside the EU/EEA?

**A1:** You must apply through the TUMonline portal and provide Preliminary Documentation (VPD), with your documents pre-evaluated through uni-assist for the Management program.

**Q2:** Where are the main locations for the Teaching at Academic Secondary Schools – Scientific Education Master of Education (M.Ed.) program offered by TUM?

**A2:** The main locations for this program are Munich, Garching, and Weihenstephan (Freising).

# Research Question

**RQ 1:** Would a multi-query formulation system improve the performance?

**RQ 2:** Would an optimization approaches, such as ensemble retriever in combination with a child-parent chunking improve the performance of the passage retriever?

**RQ 3:** Would few-shot learning enhance the performance of the system as compared to the Zero-Shot of the system?

**RQ 4:** How does the performance change when using an open-source model compared to a paid closed source model? How can open-sourced models be optimized?

# Challenge & Solution

## Challenge:

Data is valuable

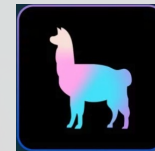


API call costs

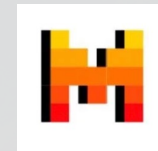


Sensitive Data should not go to a third party

## Solution:



Llama 2




Mistral




Hugging Face LLM

# Challenge & Solution


## Solution:






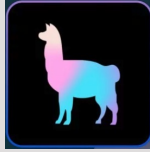



Llama 2



Mistral



Hugging Face LLM

Closed	Open		
			
GPT 3.5	Llama 2 (7B)	Mistral (7B)	Llama 2 (13B)
			
GPT 4	Mini Orca (7B)	Vicuna (7B)	

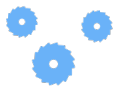


# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:



Is that right?



## Generating Questions



```
I will provide to you some source material from a section of a studyprogram. I want you to generate a question-answer pair from the source material. Please keep it concise.
```

```
### Here is the source material from the section '{section}' from the studyprogram '{studyprogram}':  
{{source}}
```

```
### Output Instruction:  
{{format_instructions}}
```

Generate 200 Question-Answer-Pairs!

# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:



Is that right?



What is the target audience for the bachelor's degree course in Teaching at Academic Secondary Schools?

What is the contact information for the Student Advisory Office in the Department of Sport and Health Sciences at Campus D for the Gesundheitswissenschaft Bachelor of Science program?

### Retrieval Quality

### Generation Quality

```
{
  "Biogenic Materials Bachelor of Science (B.Sc.): {
    "level": "Bachelor of Science (B.Sc.)",
    "studiengang": "Biogenic Materials",
    "description": "The finiteness of petroleum and fossil-based raw materials and structural changes towards a bioeconomy require the development of innovative biogenic materials. A particular challenge is combining the requirements regarding processability and high-performance properties with sustainability aspects.",
    "school": "The finiteness of petroleum and fossil-based raw materials and structural changes towards a bioeconomy require the development of innovative biogenic materials. A particular challenge is combining the requirements regarding processability and high-performance properties with sustainability aspects.",
    "school_website": "https://www.cs.tum.de/studieninteressierte/studiengaenge/biogene-werkstoffe/",
    "Type of Study": "Full Time",
    "Standard Duration of Studies": "6 semesters (fulltime)",
    "Credits": "180 ECTS",
    "Main Locations": "Straubing",
    "Application Period": "Winter semester: 15.05. – 15.07.",
    "Admission Category": "Unrestricted Admission",
    "Start of Degree Program": "Winter Semester (October)",
    "Costs": "Student Fees: 72.00 €, Tuition fees for international students",
    "Required Language Proficiency": "German",
    "Program profile": "In the light of the dramatic environmental and climate changes in recent years and the finiteness of fossil raw materials like petroleum, material sciences specializing in biogenic materials is seen as a progressive and visionary discipline. The course "Biogenic Materials" provides specific knowledge about biodegradable, eco-friendly, or environmental damage-reducing materials obtained from renewable resources. The course also creates awareness of sustainable use of resources among the future experts in material science. The technical details of extraction, transformation, shaping and processing, possible areas of application and disposal as well as particular structure-property-interrelation of materials of all classes are further aspects that are covered in the course.
```

# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:

Is that right?

Retrieval Quality	Generation Quality
<pre>{   "Biogenic Materials Bachelor of Science (B.Sc.)": {     "level": "Bachelor of Science (B.Sc.)",     "studiengang": "Biogenic Materials",     "description": "The finiteness of petroleum and fossil-based raw materials and structural changes towards a bioeconomy require the development of innovative biogenic materials. A particular challenge is combining the requirements regarding processability and high-performance properties with sustainability aspects.",     "school": "The finiteness of petroleum and fossil-based raw materials and structural changes towards a bioeconomy require the development of innovative biogenic materials. A particular challenge is combining the requirements regarding processability and high-performance properties with sustainability aspects.",     "school_website": "https://www.cs.tum.de/studieninteressierte/studiengaenge/biogene-werkstoffe/",     "Type of Study": "Full Time",     "Standard Duration of Studies": "6 semesters (fulltime)",     "Credits": "180 ECTS",     "Main Locations": "Straubing",     "Application Period": "Winter semester: 15.05. – 15.07.",     "Admission Category": "Unrestricted Admission",     "Start of Degree Program": "Winter Semester (October)",     "Costs": "Student Fees: 72.00 €, Tuition fees for international students",     "Required Language Proficiency": "German",     "Program profile": "In the light of the dramatic environmental and climate changes in recent years and the finiteness of fossil raw materials like petroleum, material sciences specializing in biogenic materials is seen as a progressive and visionary discipline. The course "Biogenic Materials" provides specific knowledge about biodegradable, eco-friendly, or environmental damage-reducing materials obtained from renewable resources. The course also creates awareness of sustainable use of resources among the future experts in material science. The technical details of extraction, transformation, shaping and processing, possible areas of application and disposal as well as particular structure-property interrelation of materials of all classes are further aspects that are covered in the course</pre>	

# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:

Is that right?

↓

### Retrieval Quality

### Generation Quality

Table 6.1.: Hit Rate over all possible RAG frameworks

Model	#P		er	cpr	icl	icl-er	mq-er	mq-cpr	mq-cpr-icl	mq-icl-er
Llama 2	7B	de	8.64	8.64	8.64	7.41	<b>13.58</b>	<b>13.58</b>	<b>13.58</b>	<b>13.58</b>
		en	43.21	43.21	43.21	43.21	<b>53.09</b>	<b>53.09</b>	<b>53.09</b>	<b>53.09</b>
Llama 2	13B	de	28.40	28.40	28.40	28.40	<b>34.57</b>	<b>34.57</b>	<b>34.57</b>	<b>34.57</b>
		en	50.62	50.62	50.62	50.62	<b>55.56</b>	<b>55.56</b>	<b>55.56</b>	<b>55.56</b>
GPT 3.5		de	56.79	56.79	58.02	55.56	66.67	64.20	<b>70.37</b>	65.43
		en	44.44	41.98	41.98	43.21	<b>46.91</b>	44.44	45.68	45.68
GPT 4		de	61.73	61.73	61.73	61.73	69.14	<b>66.67</b>	65.43	<b>66.67</b>
		en	65.43	66.67	66.67	66.67	<b>75.31</b>	72.84	72.84	72.84
Mistral	7B	de	39.51	39.51	39.51	39.51	48.15	<b>51.85</b>	<b>51.85</b>	49.38
		en	51.85	51.85	53.09	51.85	56.79	56.79	56.79	56.79
Orca Mini	7B	en	44.44	44.44	44.44	44.44	50.62	49.38	<b>51.85</b>	<b>51.85</b>
Vicuna	7B	en	1.23	1.23	1.23	1.23	3.70	3.70	3.70	<b>6.17</b>

er = Ensemble Retriever, cpr = Child-Parent-Retriever, icl = In-Context-Learning, mq = Multi-Query.



# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:

Is that right?

↓

Retrieval Quality

	bm25_0.5
Llama2_7B_de	8.64

Generation Quality

**Correct context**  
=  
**Good context**

**Incorrect context**  
=  
**Bad context**

# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:

Is that right?

↓




# Testing & Evaluation

## Self-Evaluation:


**2**

" If we want to study <x> we have the following conditions:



Is that right?

↓



### Retrieval Quality

### Generation Quality

Table 5.1.: Model Evaluation Summary

Match	Question	True Answer	Predicted Answer	Rel.	Coh.	Flu.	Faith.
1	Question 1	True Answer 1	Answer 1	5	5	5	5
1	Question 2	True Answer 2	Answer 2	3	4	4	2
1	Question 3	True Answer 3	Answer 3	5	5	5	4
<i>(Entries 4-10 omitted)</i>							
0	Question 11	True Answer 11	Answer 11	1	1	5	1
0	Question 12	True Answer 12	Answer 12	3	1	3	2
0	Question 13	True Answer 13	Answer 13	5	5	5	2
<i>(Entries 14-20 omitted)</i>							




# Testing & Evaluation

## Self-Evaluation:


2

" If we want to study <x> we have the following conditions:



Is that right?

↓



### Retrieval Quality

Table 5.1.: Model Evaluation Summary

Match	Question	True Answer	Predicted Answer	Rel.	Coh.	Flu.	Faith.
1	Question 1	True Answer 1	Answer 1	5	5	5	5
1	Question 2	True Answer 2	Answer 2	3	4	4	2
1	Question 3	True Answer 3	Answer 3	5	5	5	4
<i>(Entries 4-10 omitted)</i>							
0	Question 11	True Answer 11	Answer 11	1	1	5	1
0	Question 12	True Answer 12	Answer 12	3	1	3	2
0	Question 13	True Answer 13	Answer 13	5	5	5	2
<i>(Entries 14-20 omitted)</i>							

### Generation Quality

Table 5.2.: RAG Confusion Matrix, Metric: Relevance, Threshold: 5


	Correct	Incorrect
Match	7	3
No Match	4	6

# Testing & Evaluation

## Self-Evaluation:


2


" If we want to study <x> we have the following conditions:



Is that right?

↓





Retrieval Quality

Generation Quality

**Evaluation of one Dataset**


identified_keys	match	model_name	language	modul_iteration	Question	True_Answer	Predicted_Answer	Relevance_Score
22	1	Llama2_7B	de	icl	Welche Sprachnachweise sind für die Bewerbung erforderlich?	Bei der Bewerbung müssen ausreichende Deutschkenntnisse nachgewiesen werden.	For the application to the study program "Biology and Chemistry", sufficient proof of language skills is required.	1
104	1	Llama2_7B	de	icl	Wann ist der Semesterstart für das Studienprogramm?	Nur zum Wintersemester möglich	The semester start for the "AgriFood Economics" study program is in October.	4
35	1	Llama2_7B	de	icl	Welche Masterstudiengänge werden im Studienprogramm angeboten?	Im Studienprogramm werden die Masterstudiengänge Maschinenbau und Materialwissenschaften angeboten.	In the study program "Forest Science and Resource Management", the Master's programs "Sustainable Management of Bioresources" and "Forest and Nature Conservation" are offered.	2
88	0	Llama2_7B	de	icl	An wen kann ich mich bei Fragen zum Wirtschaftsinformatik Studium wenden?	Du kannst dich an die Studiengangskordinatorin Frau Schmidt wenden.	For questions regarding the Economic Informatics program, you can contact the program coordinator Ms. Schmidt.	4
101	0	Llama2_7B	de	icl	Wie viele ECTS-Punkte umfasst das Studienprogramm?	Das Studienprogramm umfasst 180 ECTS-Punkte.	The "Technology of Biogenic Resources" Bachelor's program comprises 180 ECTS points.	5
181	0	Llama2_7B	de	icl	Welche Art von Eignungsverfahren wird für den Masterstudiengang durchgeführt?	Es wird ein Eignungsverfahren für Masterstudiengänge in Biologie und Chemie durchgeführt.	For the Master's program in Agricultural Biology, an aptitude test is conducted.	4

# Testing & Evaluation

## Self-Evaluation:


2


" If we want to study <x> we have the following conditions:



Is that right?

↓





Retrieval Quality

Generation Quality

**Evaluation of one Dataset**

```

RELEVANCY_SCORE_CRITERIA = """
Relevance(1-5) - degree to which the answer covers the content in the reference answer. \
The answer should include only the information relevant to the question and any information other than the ones \
mentioned in the reference answer should be penalized. \
Please provide an explanation for your rating of the reference answer and the rating itself on a scale of 0 (bad \
quality) to 5 (perfect quality) as integer number.
"""

RELEVANCY_SCORE_STEPS = """
1. Read the question, generated answer and the reference answer carefully.
2. Compare the information in the generated answer to the reference answers and check if all points in it, are \
relevant to the question.
3. Assess how well the answer covers the main query of the question, and how much irrelevant or redundant \
information it contains.
4. Assign a relevance score from 1 to 5.
"""


```

# Testing & Evaluation

## Self-Evaluation:


2


" If we want to study <x> we have the following conditions:



Is that right?

↓





Retrieval Quality

Generation Quality

**Evaluation of one Dataset**

```

EVALUATION_USER_PROMPT_TEMPLATE = ""
Example:
Question:
{{question}}
Generated Answer:
{{gen_answer}}
Reference Answer:
{{ref_answer}}
Evaluation Form:
{{metric_name}}
Please provide your output in two parts separate as a python dictionary with keys rating and explanation.

First the rating in an integer followed by the explanation of the rating, in dictionary:

rating:<your rating>, explanation:<your explanation>
```

# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:

Is that right?

↓



### Retrieval Quality Evaluation of one Dataset

### Generation Quality

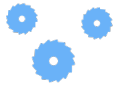
identified_keys	match	model_name	language	modul_iteration	Question	True_Answer	Predicted_Answer	Relevance_Score
22	1	Llama2_7B	de	icl	Welche Sprachnachweise sind für die Bewerbung erforderlich?	Bei der Bewerbung müssen ausreichende Deutschkenntnisse nachgewiesen werden.	For the application to the study program "Biology and Chemistry", sufficient proof of language skills is required.	1
104	1	Llama2_7B	de	icl	Wann ist der Semesterstart für das Studienprogramm?	Nur zum Wintersemester möglich	The semester start for the "AgriFood Economics" study program is in October.	4
35	1	Llama2_7B	de	icl	Welche Masterstudiengänge werden im Studienprogramm angeboten?	Im Studienprogramm werden die Masterstudiengänge Maschinenbau und Materialwissenschaften angeboten.	In the study program "Forest Science and Resource Management", the Master's programs "Sustainable Management of Bioresources" and "Forest and Nature Conservation" are offered.	2
88	0	Llama2_7B	de	icl	An wen kann ich mich bei Fragen zum Wirtschaftsinformatik Studium wenden?	Du kannst dich an die Studiengangskordinatorin Frau Schmidt wenden.	For questions regarding the Economic Informatics program, you can contact the program coordinator Ms. Schmidt.	4
101	0	Llama2_7B	de	icl	Wie viele ECTS-Punkte umfasst das Studienprogramm?	Das Studienprogramm umfasst 180 ECTS-Punkte.	The "Technology of Biogenic Resources" Bachelor's program comprises 180 ECTS points.	5
181	0	Llama2_7B	de	icl	Welche Art von Eignungsverfahren wird für den Masterstudiengang durchgeführt?	Es wird ein Eignungsverfahren für Masterstudiengänge in Biologie und Chemie durchgeführt.	For the Master's program in Agricultural Biology, an aptitude test is conducted.	4

# Testing & Evaluation

## Self-Evaluation:


**2**

" If we want to study <x> we have the following conditions:



Is that right?

↓



### Retrieval Quality

### Generation Quality

Table 6.2.: Metric: Relevance, Threshold: 5; Top 3 TP: Correct Match & Correct Response

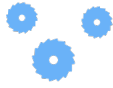
Match	Llama2 7B (en-mq-er)		Mistral 7B (en-mq-er)		Mistral 7B (de-icl-er)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	9(TP)	1(FN)	9(TP)	1(FN)	9(TP)	1(FN)
0	2(FP)	8(TN)	5(FP)	5(TN)	4(FP)	6(TN)

# Testing & Evaluation

## Self-Evaluation:


2

" If we want to study <x> we have the following conditions:



Is that right?

↓



### Retrieval Quality

### Generation Quality

Table 6.4.: Metric: Faithfulness, Threshold: 5; Top 3 TP: Correct Match & Correct Response

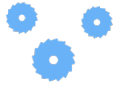
Match	ChatGPT4 (en-mq-er)		Llama2 7B (en-er)		Mistral 7B (en-mq-er)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	10(TP)	0(FN)	9(TP)	1(FN)	9(TP)	1(FN)
0	5(FP)	5(TN)	4(FP)	6(TN)	5(FP)	5(TN)

# Testing & Evaluation

## Self-Evaluation:

2

“ If we want to study <x> we have the following conditions:



Is that right?

↓




Table 6.5.: Metric: Relevance, Threshold: 5

Match	ChatGPT4 (en-mq-er)		Llama2 7B (en-er)		Mistral 7B (en-mq-er)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	7(TP)	3(FN)	7(TP)	3(FN)	9(TP)	1(FN)
0	4(FP)	6(TN)	4(FP)	6(TN)	5(FP)	5(TN)

Table 6.6.: Metric: Coherence, Threshold: 5

Match	ChatGPT4 (en-mq-er)		Llama2 7B (en-er)		Mistral 7B (en-mq-er)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	10(TP)	0(FN)	8(TP)	2(FN)	9(TP)	1(FN)
0	3(FP)	7(TN)	5(FP)	5(TN)	3(FP)	7(TN)

Table 6.7.: Metric: Fluency, Threshold: 5

Match	ChatGPT4 (en-mq-er)		Llama2 7B (en-er)		Mistral 7B (en-mq-er)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	10(TP)	0(FN)	8(TP)	2(FN)	10(TP)	0(FN)
0	7(FP)	3(TN)	9(FP)	1(TN)	6(FP)	4(TN)




# Testing & Evaluation

## Self-Evaluation:


**2**

" If we want to study <x> we have the following conditions:



Is that right?

↓



### Retrieval Quality

### Generation Quality

Table 6.8.: Metric: Faithfulness, Threshold: 4

Match	ChatGPT4 (en-mq-er)		ChatGPT4 (de-mq-er)		Llama2 7B (en-er)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	10(TP)	0(FN)	10(TP)	0(FN)	9(TP)	1(FN)
0	5(FP)	5(TN)	6(FP)	4(TN)	7(FP)	3(TN)

# Testing & Evaluation

## Self-Evaluation:

**2**

" If we want to study <x> we have the following conditions:

Is that right?

↓

		Retrieval Quality			Generation Quality			
Table 6.10.: Continuation of table 6.9; Metric: Faithfulness, Average Values								
<b>Mutli-Query &amp; Ensemble Retriever</b>								
		Llama 2 7B	Llama 2 13B	GPT 3.5	GPT 4	Mistral 7B	Orca Mini 7B	Vicuna 7B
de	Match	<b>4.0</b>	3.0	3.9	<b>4.8</b>	<b>4.4</b>	<i>x</i>	<i>x</i>
	No Match	1.9	2.0	2.7	3.4	3.3	<i>x</i>	<i>x</i>
en	Match	<b>4.5</b>	3.8	3.6	<b>5.0</b>	<b>4.6</b>	<b>4.0</b>	1.4
	No Match	2.6	2.7	1.8	3.3	3.5	2.9	0.3
<b>Multi-Query &amp; Child-Parent-Retriever</b>								
		Llama 2 7B	Llama 2 13B	GPT 3.5	GPT 4	Mistral 7B	Orca Mini 7B	Vicuna 7B
de	Match	3.6	2.8	3.2	2.5	<b>4.0</b>	<i>x</i>	<i>x</i>
	No Match	2.0	3.3	2.6	3.1	3.8	<i>x</i>	<i>x</i>
en	Match	3.2	2.7	3.4	3.6	2.3	3.4	0.7
	No Match	2.4	2.1	1.5	2.2	2.7	3.2	0.5

# Testing & Evaluation

## Self-Evaluation:

2

" If we want to study <x> we have the following conditions:

Is that right?

↓

### Retrieval Quality

### Generation Quality

#### Multi-Query & Child-Parent-Retriever & In-Context-Learning

		Llama 2 7B	Llama 2 13B	GPT 3.5	GPT 4	Mistral 7B	Orca Mini 7B	Vicuna 7B
de	Match	2.9	2.7	3.6	3.3	2.6	<i>x</i>	<i>x</i>
	No Match	3.1	2.8	2.9	3.1	4.5	<i>x</i>	<i>x</i>
en	Match	3.6	3.2	3.1	3.5	3.2	2.6	0.3
	No Match	3.2	2.3	2.5	2.7	2.2	3.2	1.0

#### Multi-Query & In-Context-Learning & Ensemble Retriever

		Llama 2 7B	Llama 2 13B	GPT 3.5	GPT 4	Mistral 7B	Orca Mini 7B	Vicuna 7B
de	Match	3.6	3.3	3.9	<b>4.0</b>	<b>4.1</b>	<i>x</i>	<i>x</i>
	No Match	3.1	2.9	2.1	3.2	<b>4.3</b>	<i>x</i>	<i>x</i>
en	Match	3.1	<b>4.6</b>	<b>4.0</b>	3.8	<b>4.2</b>	3.2	2.1
	No Match	2.5	2.7	1.9	2.9	2.3	2.9	0.5

# Optimization

Table 6.1.: Hit Rate over all possible RAG frameworks

Model	#P		er	cpr	icl	icl- er	mq- er	mq- cpr	mq- cpr-icl	mq- icl- er
Llama 2	7B	de	8.64	8.64	8.64	7.41	<b>13.58</b>	<b>13.58</b>	<b>13.58</b>	<b>13.58</b>
		en	43.21	43.21	43.21	43.21	<b>53.09</b>	<b>53.09</b>	<b>53.09</b>	<b>53.09</b>
Llama 2	13B	de	28.40	28.40	28.40	28.40	<b>34.57</b>	<b>34.57</b>	<b>34.57</b>	<b>34.57</b>
		en	50.62	50.62	50.62	50.62	<b>55.56</b>	<b>55.56</b>	<b>55.56</b>	<b>55.56</b>
GPT 3.5		de	56.79	56.79	58.02	55.56	66.67	64.20	<b>70.37</b>	65.43
		en	44.44	41.98	41.98	43.21	<b>46.91</b>	44.44	45.68	45.68
GPT 4		de	61.73	61.73	61.73	61.73	69.14	<b>66.67</b>	65.43	<b>66.67</b>
		en	65.43	66.67	66.67	66.67	<b>75.31</b>	<b>72.84</b>	72.84	72.84
Mistral	7B	de	39.51	39.51	39.51	39.51	48.15	<b>51.85</b>	<b>51.85</b>	49.38
		en	51.85	51.85	53.09	51.85	56.79	<b>56.79</b>	56.79	56.79
Orca Mini	7B	en	44.44	44.44	44.44	44.44	50.62	49.38	<b>51.85</b>	<b>51.85</b>
Vicuna	7B	en	1.23	1.23	1.23	1.23	3.70	3.70	3.70	<b>6.17</b>

er = Ensemble Retriever, cpr = Child-Parent-Retriever, icl = In-Context-Learning, mq = Multi-Query.

# Optimization

Table 6.11.: Hit Rate over all possible RAG frameworks

Model	#P		mq-icl-er	mq-icl-er-optimized
Llama 2	13B	en	55.56	<b>74.07</b>
GPT 4		en	72.84	<b>81.48</b>
Mistral	7B	en	56.79	<b>67.90</b>

# Conclusion & Future Work

- Change RAG Evaluation from Black Box to Grey Box
- Multi-Query helped a lot because of Open-Domain specific questions
  - Hypothetical Doc will help even more
  - BM25 is surprisingly strong
- Different Moduls can be beneficial. Especially Information Compression
- Lost in the middle!
- Fine-Tuning or using alternatives to cosine-similarity is also possible

<Thanks for the attention>



## **Gentrit Fazlija**

MSc Student Mathematics in Data Science

Technical University of Munich (TUM)  
TUM School of CIT  
Department of Computer Science (CS)  
Chair of Software Engineering for Business  
Information Systems (sebis)

Boltzmannstraße 3  
85748 Garching bei München

+49.89.289.  
gentrit.fazlija@tum.de  
[www.matthes.in.tum.de](http://www.matthes.in.tum.de)

