sebis

Master thesis

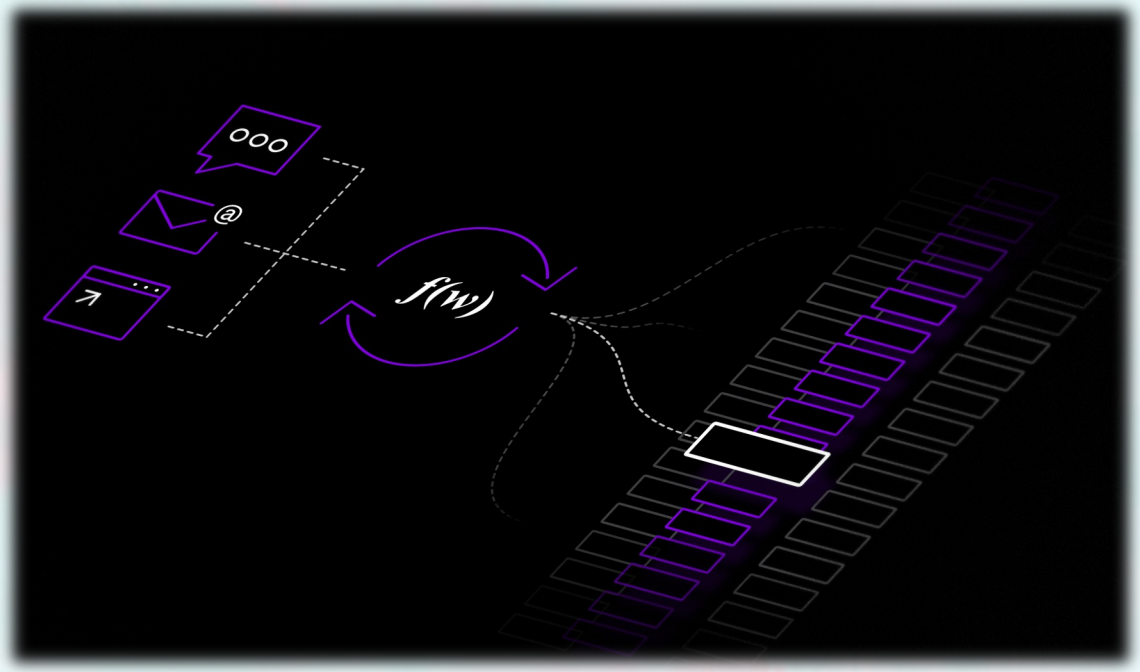
Toward Optimizing a Retrieval Augmented Generation Pipeline using Large Language Model

Student: Gentrit Fazlija

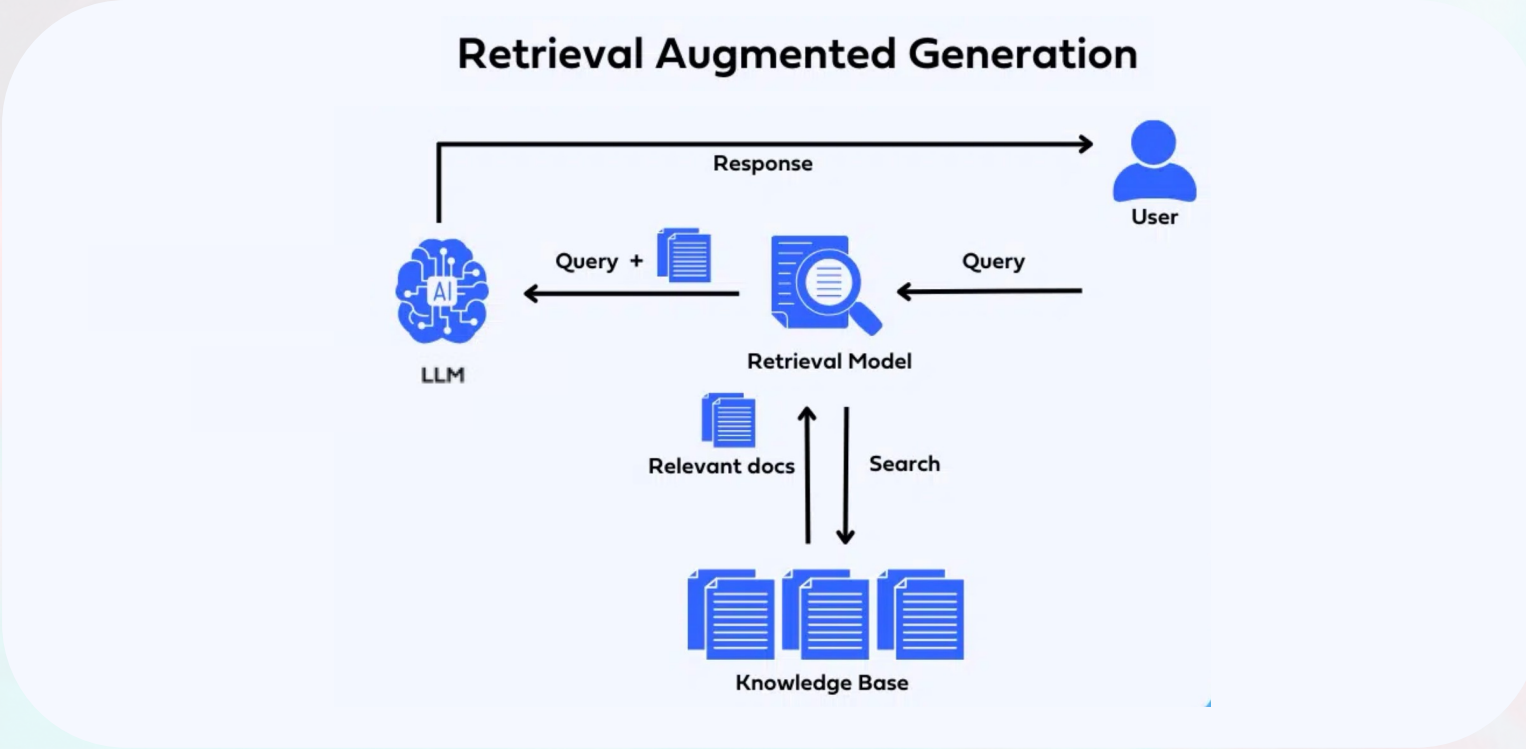
Supervisor: Anum Afzal

Professor: Prof. Dr. Florian Matthes

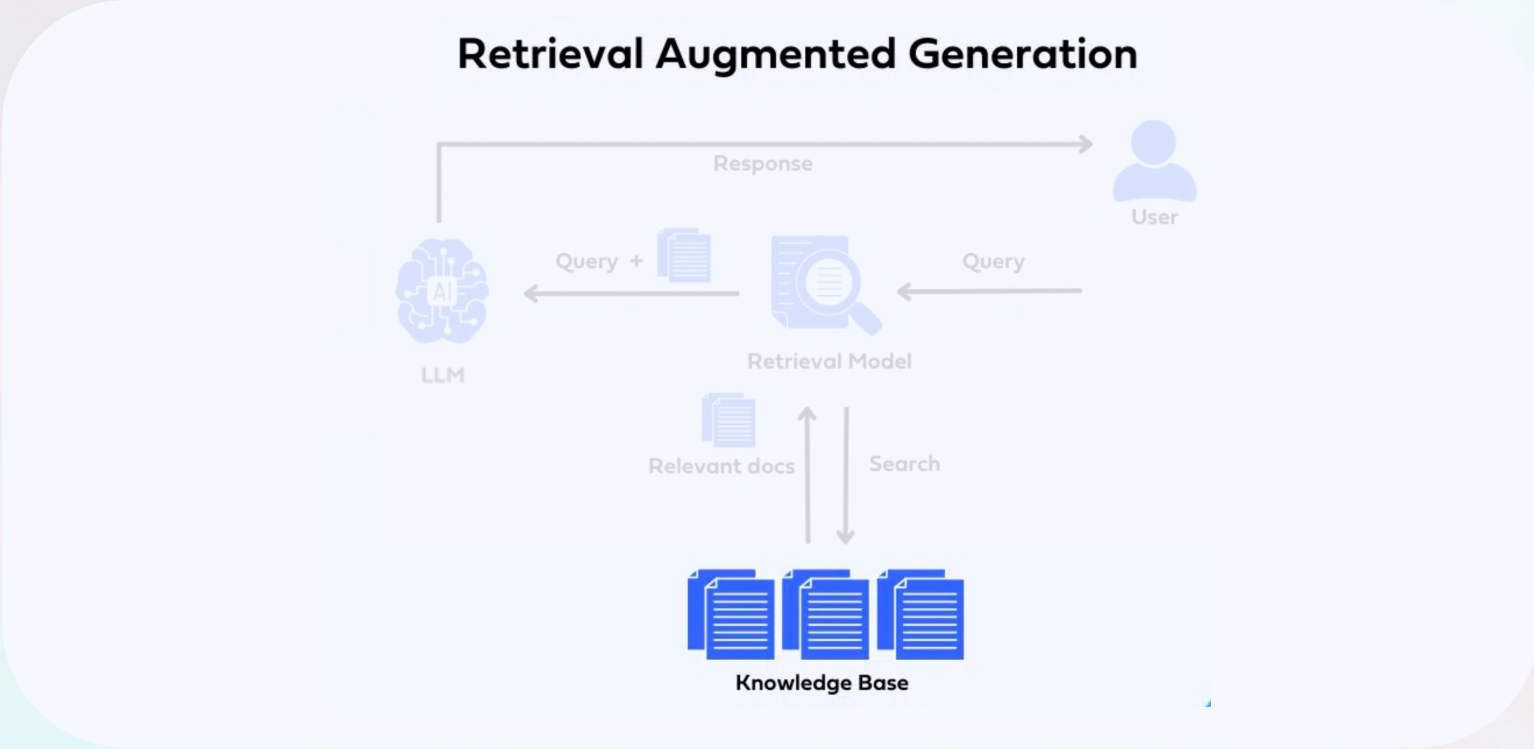
Table of content



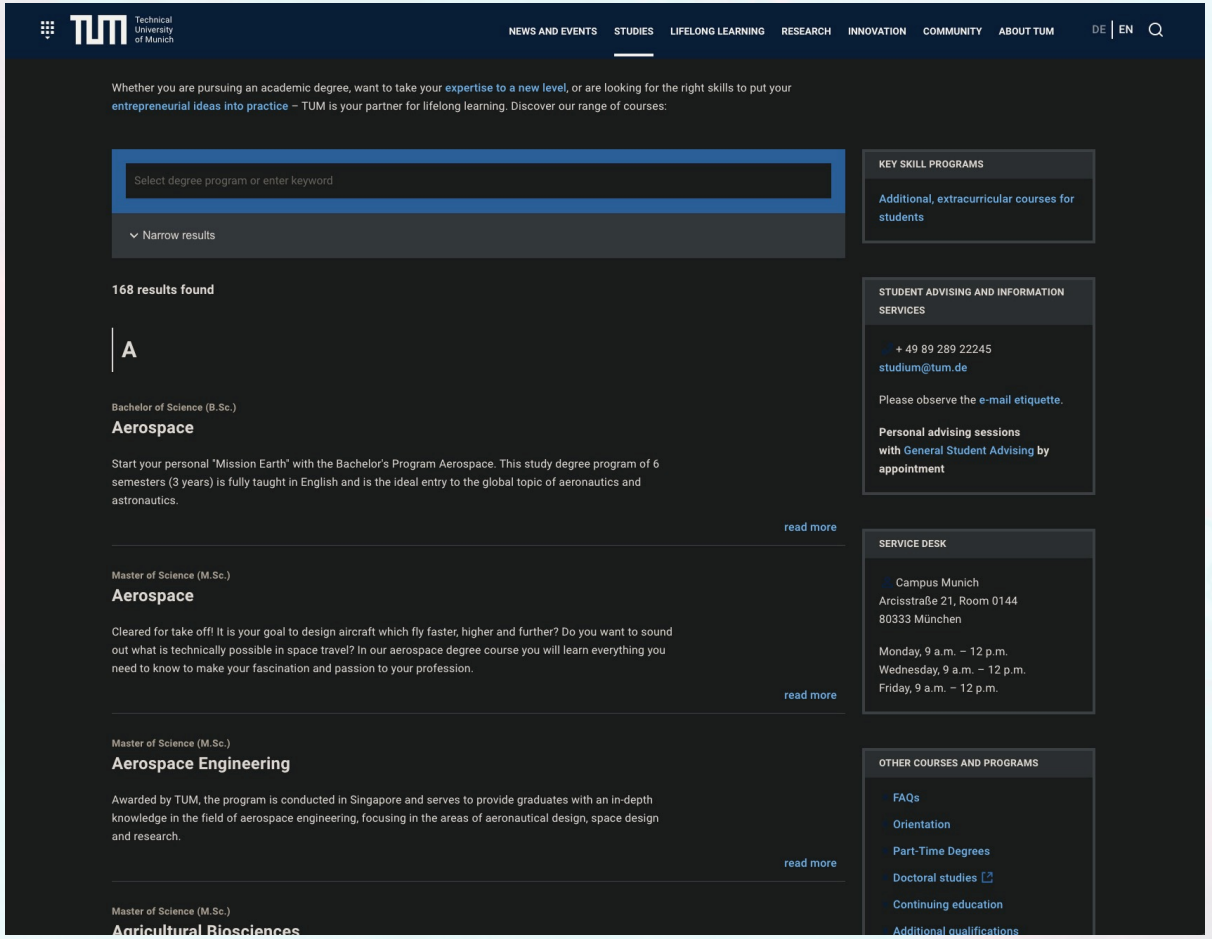
Key Components & Motivation



Key Components & Motivation



Knowledge Base

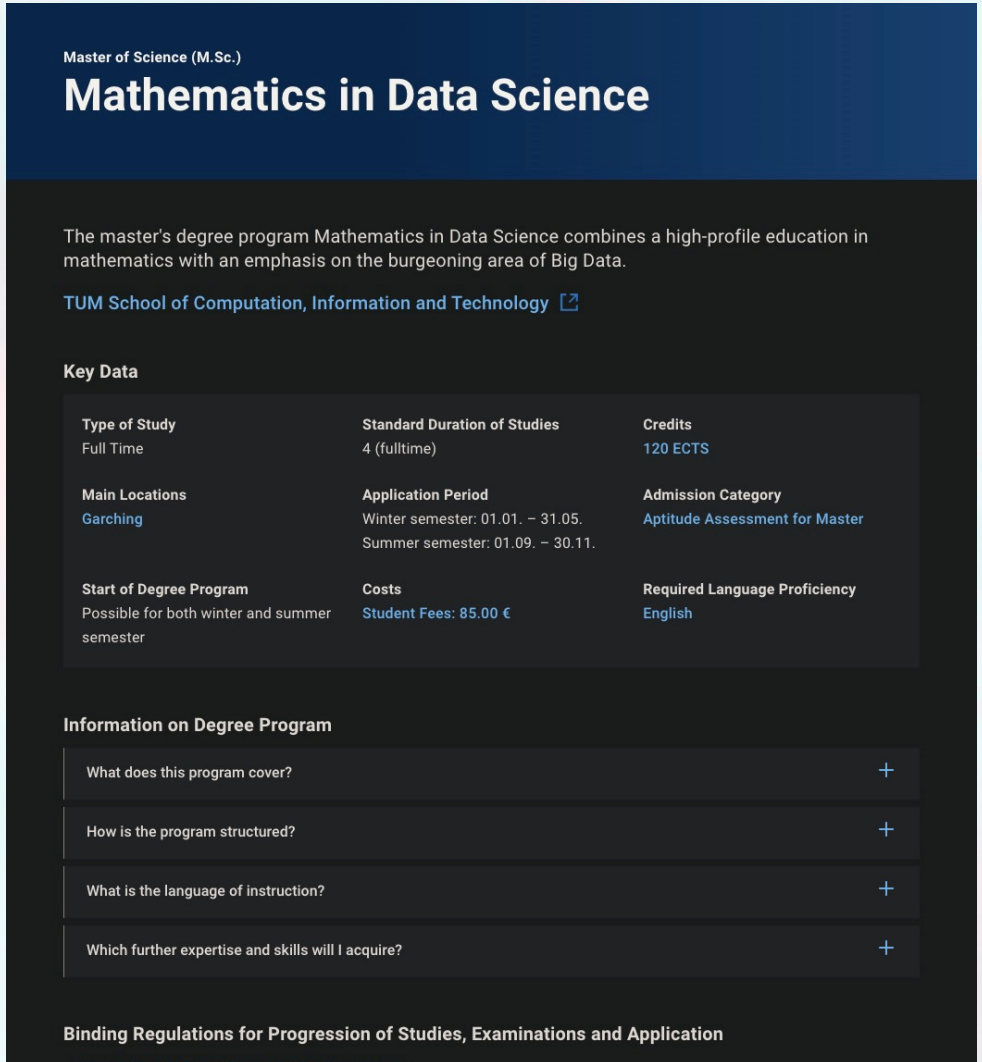


https://www.tum.de/en/studies/degree-programs

Knowledge Base

What does this program cover?

How is the program structured?



How do I enroll?

How long does the master's thesis take?

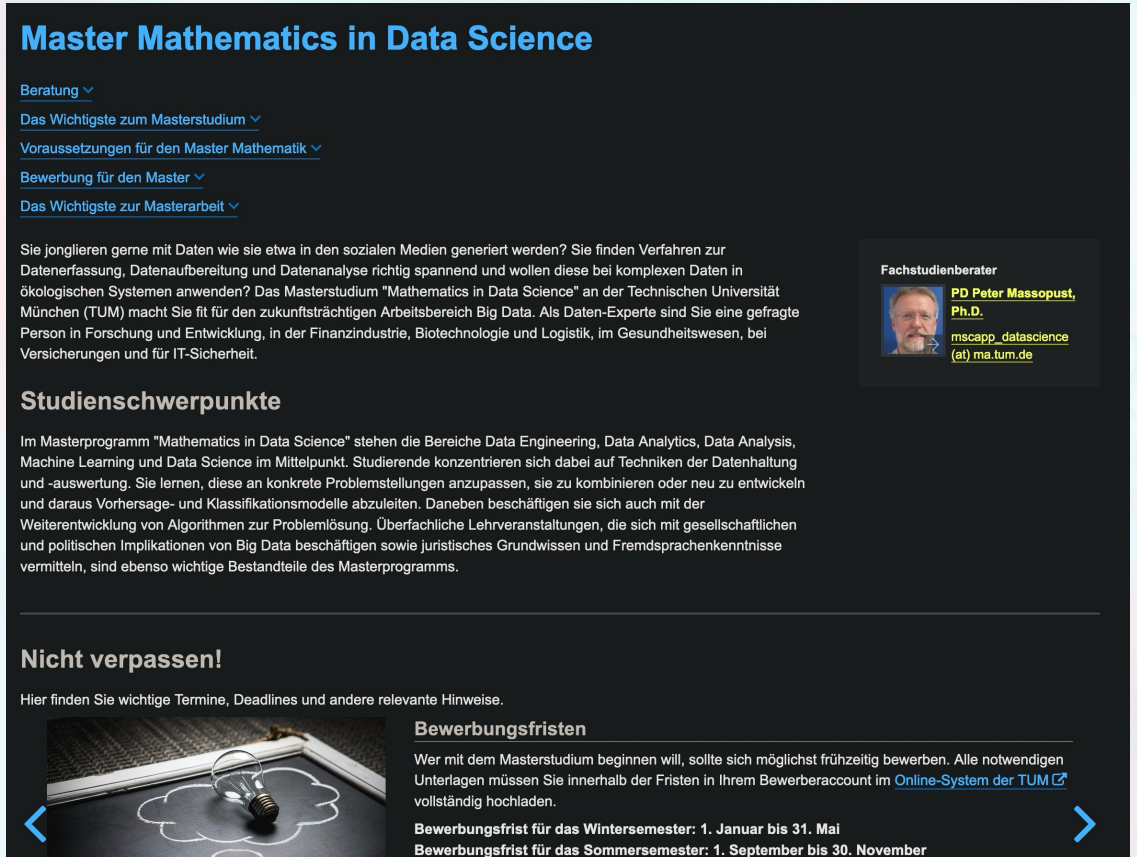
Knowledge Base

What does this program cover?

How is the program structured?

How do I enroll?

How long does the master's thesis take?



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 (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.

Knowledge Base

What does this program cover?

How is the program structured?

Fachprüfungs- und Studienordnung
für den Masterstudiengang Mathematics in Data Science
an der Technischen Universität München

Vom 17. August 2023

Aufgrund von Art. 9 Satz 1 und 2 in Verbindung mit Art. 80 Abs. 1 Satz 1, Art. 84 Abs. 2 Satz 1 sowie Art. 90 Abs. 1 Satz 2 des Bayerischen Hochschulinnovationsgesetzes (BayHIG) erlässt die Technische Universität München folgende Satzung:

Inhaltsverzeichnis:

How do I enroll?

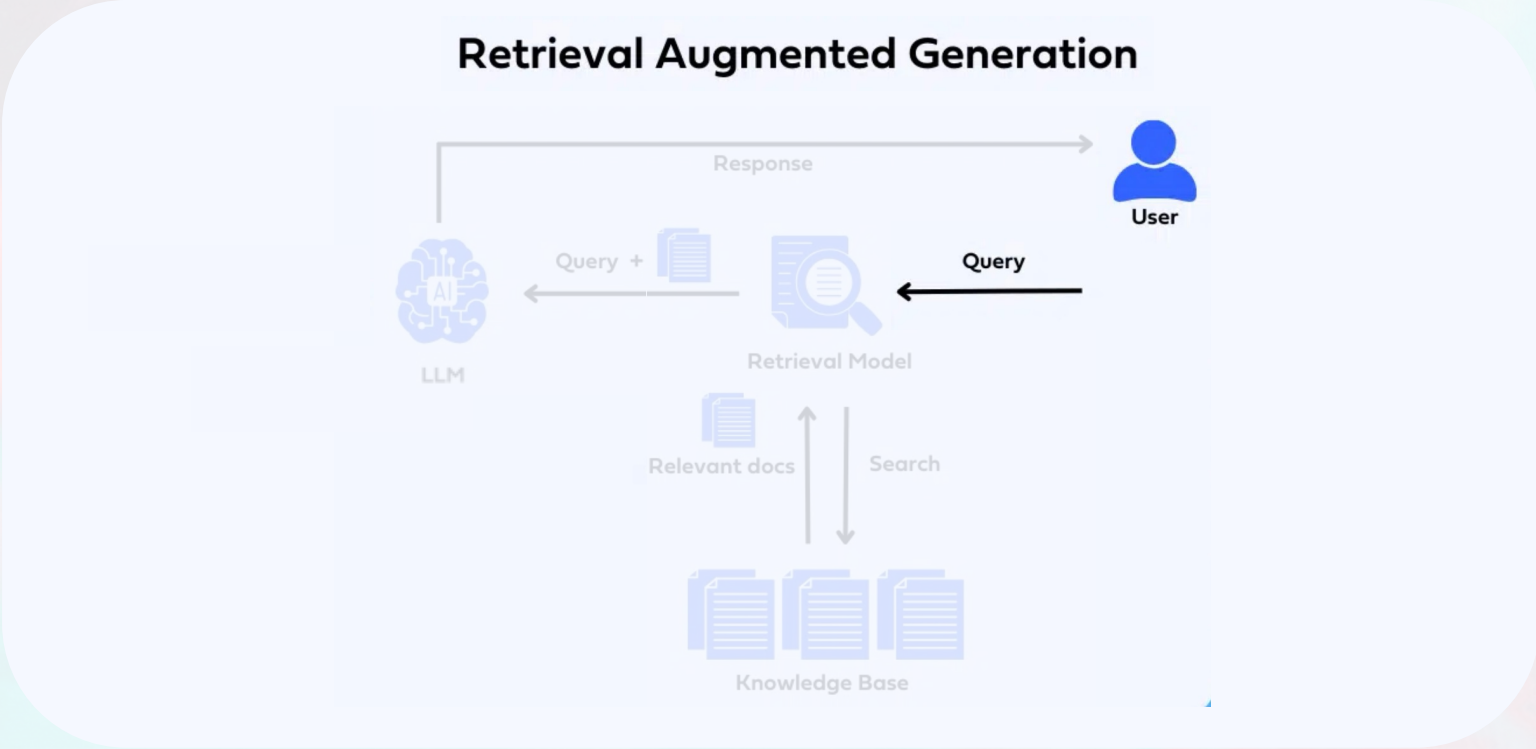
How long does the master's thesis take?

Solution

I want to create a state-of-the-art information retrieval model, capable of querying TUM course data. This will help existing students understand their own degree program better and at the same time support applicants.

TUM GPT

Key Components & Motivation



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

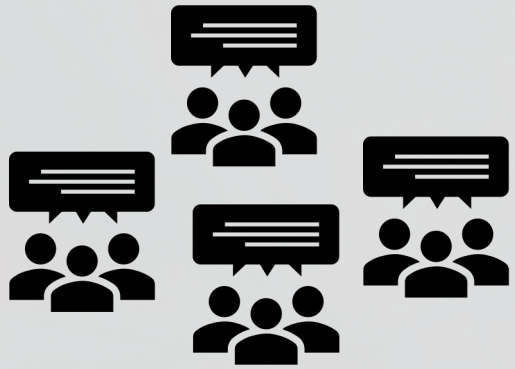
+

Students ask the same question in different manners

+

They might not even know exactly what they are looking for

Solution:



Research Question

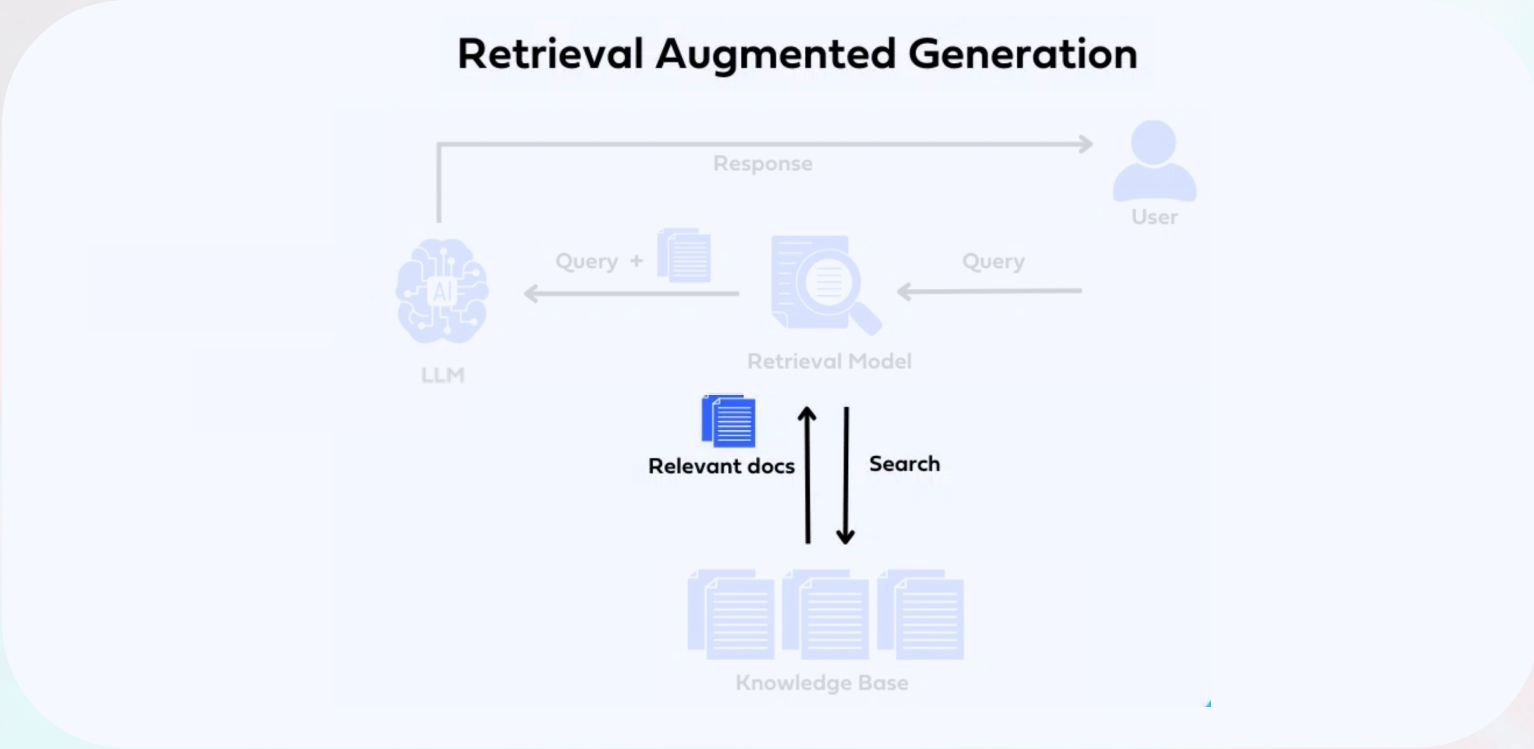
RQ 1: Would a multi-query formulation system improve the performance?

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?

Key Components & Motivation



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. [2]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) [1]Eine **Laborleistung** beinhaltet je nach Fachdisziplin Versuche, Messungen, Arbeiten im Feld, Feldübungen etc. mit dem Ziel der Durchführung, Auswertung und Erkenntnisgewinnung. [2]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. [3]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

Challenge:

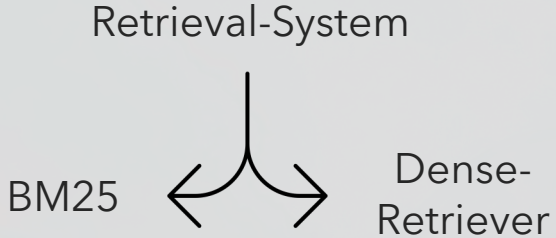
Data is very similar in between different study programs

+

Data within one study program has specific words being specific things

Solution:

2



Research Question

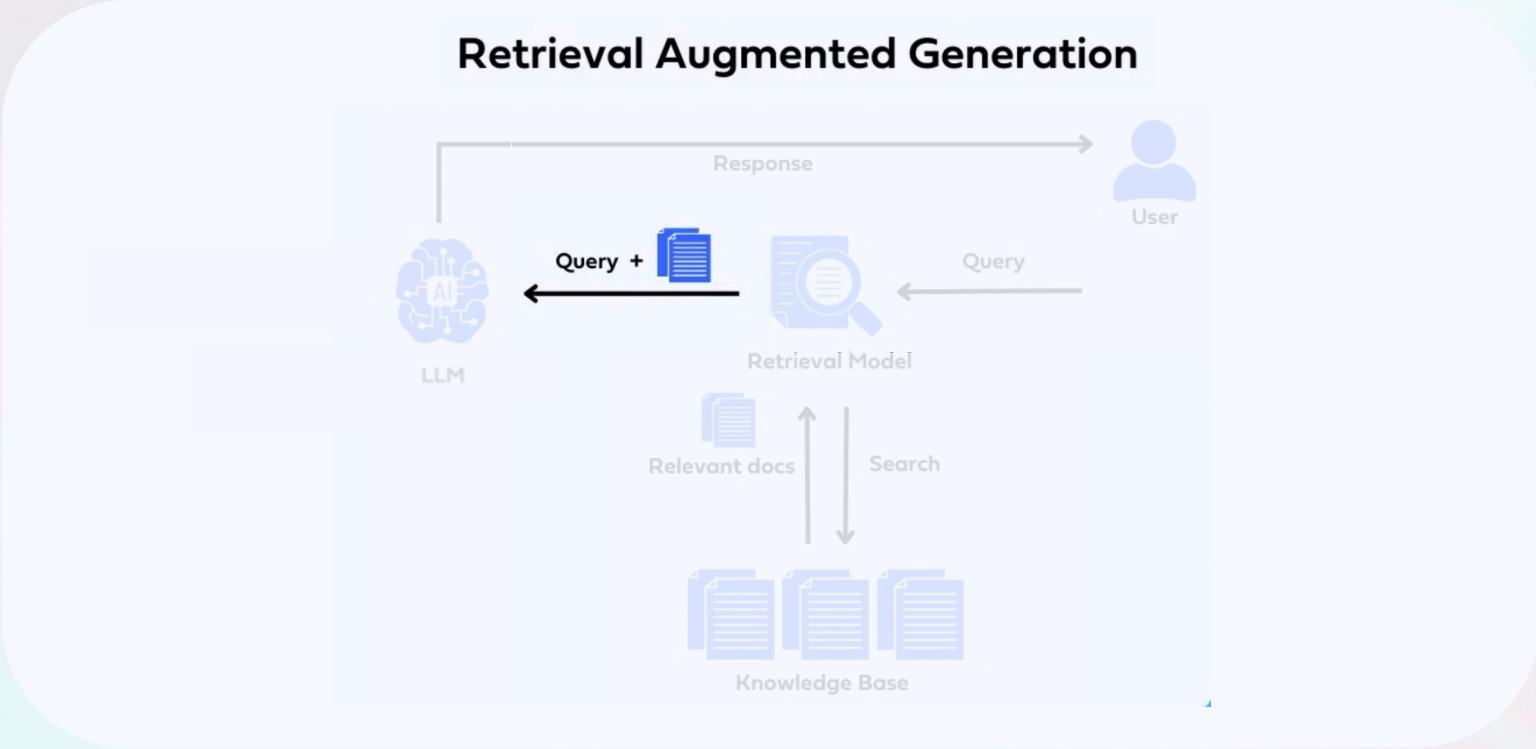
RQ 1: Would a multi-query formulation system improve the performance?

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?

Key Components & Motivation



Challenge & Solution

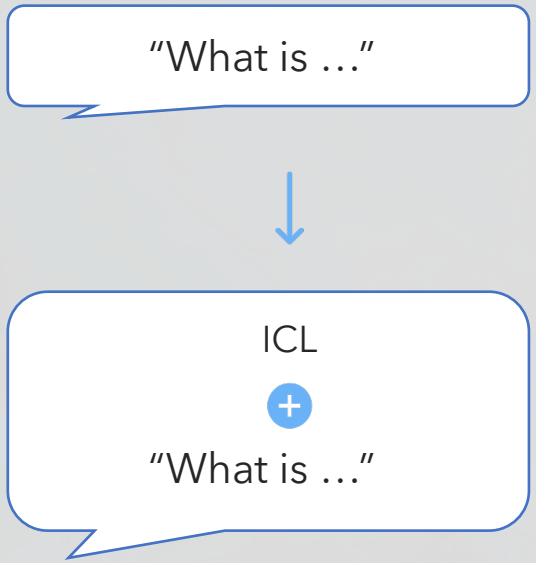
Challenge:

LLMs tend to hallucinate

+

LLMs tend to give highly divergent outputs

Solution:



Research Question

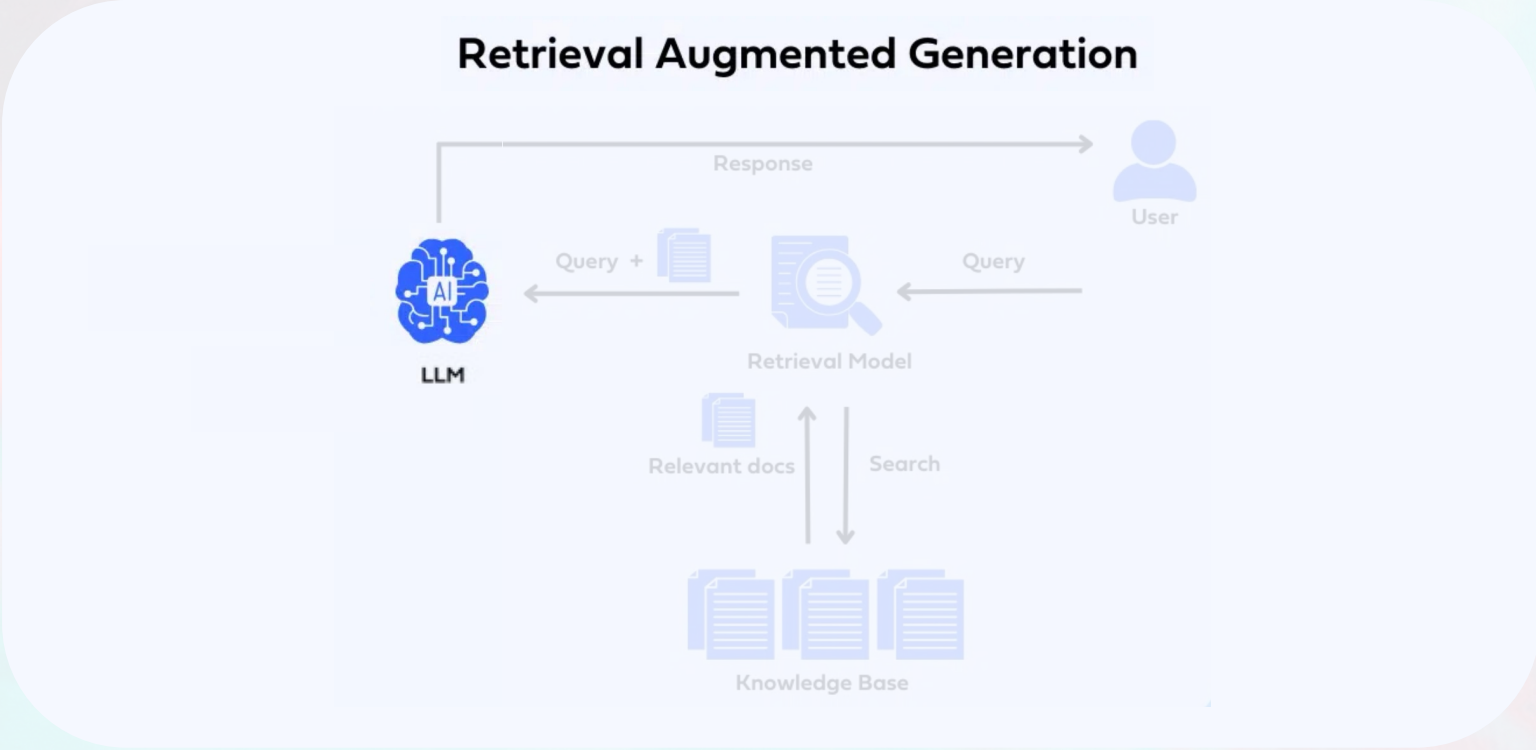
RQ 1: Would a multi-query formulation system improve the performance?

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?

Key Components & Motivation



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

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?

Testing & Evaluation

Binary Classification:

"How much time do I have for the master thesis?"

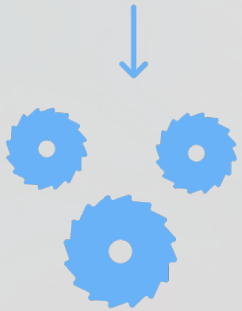
"When is the application deadline for Data Science?"

"Do I need a Proof of English language skills?"

Human Evaluation:

1

"Tell me about the Conditions I have to fulfill when I want to study <x>?"

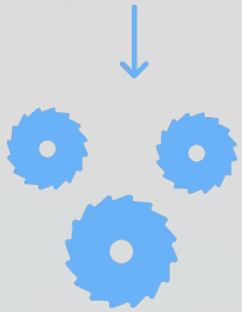


Testing & Evaluation

Human Evaluation:

1

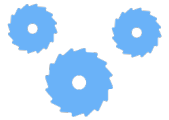
"Tell me about the Conditions I have to fulfill when I want to study <x>?"



Self-Evaluation:

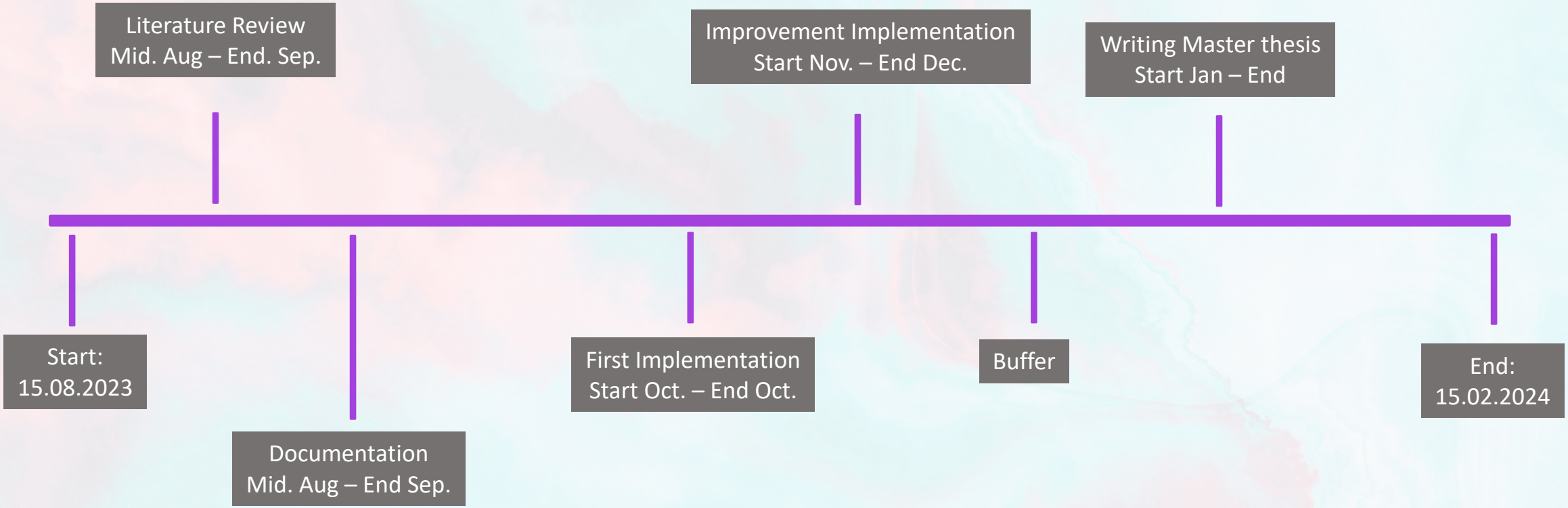
2

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



Is that right?

Outlook


Literature Review
Mid. Aug – End. Sep.
Improvement Implementation
Start Nov. – End Dec.
Writing Master thesis
Start Jan – End
Start:
15.08.2023
Documentation
Mid. Aug – End Sep.
First Implementation
Start Oct. – End Oct.
Buffer
End:
15.02.2024

<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
wwwmatthes.in.tum.de