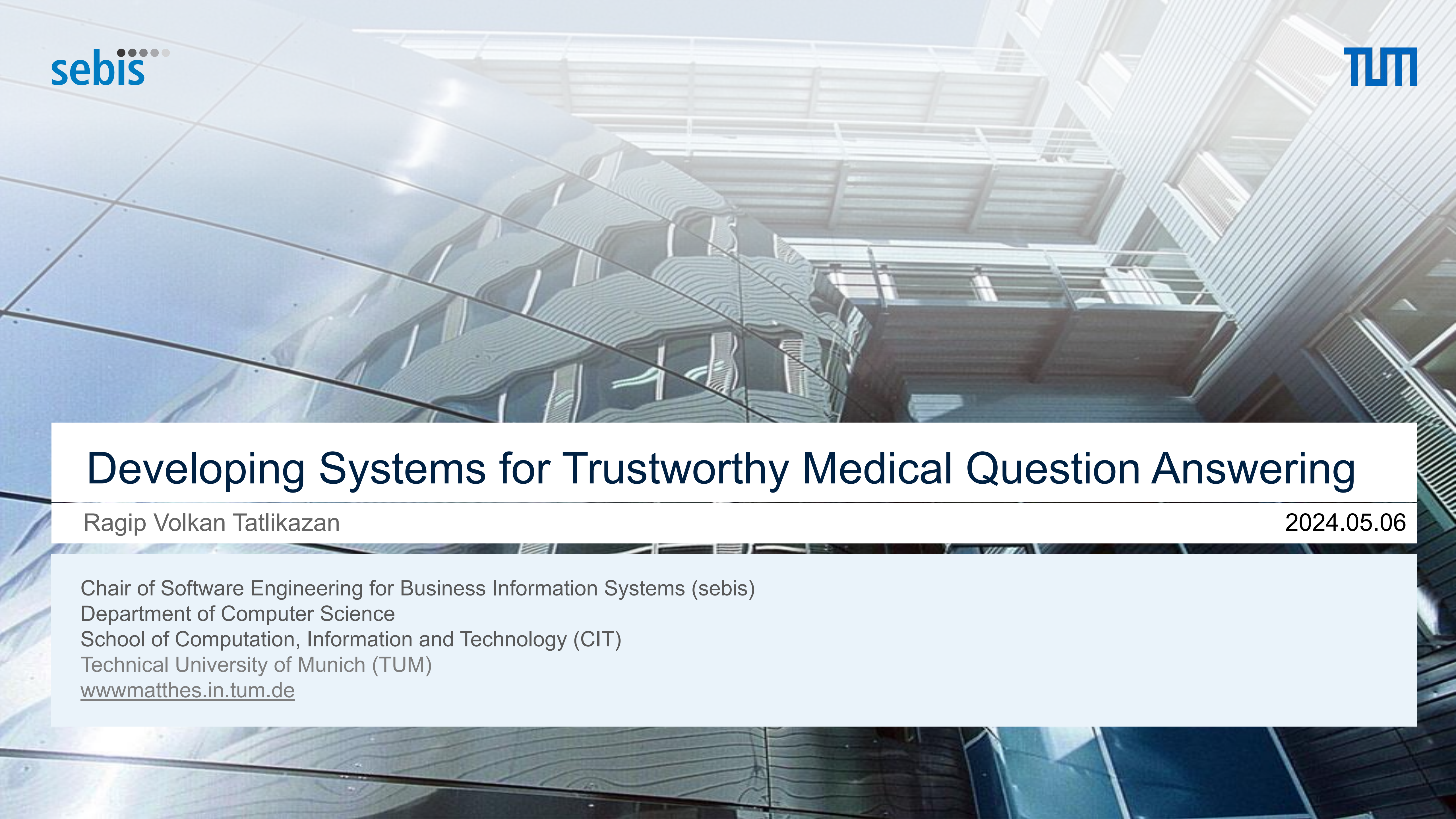


Developing Systems for Trustworthy Medical Question Answering

Ragip Volkan Tatlikazan

2024.05.06

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

- Developing Systems for Trustworthy Medical Question Answering (T-MQA)
  - Motivation
  - Research Questions
    - RQ1: Approach
    - RQ2: System
    - RQ3: Users
  - Timeline

T-MQA: Motivation

Question & answering is challenging in the **scientific domain**, in particular covering **biomedical** and **health-related** claims, which can often be **uncertain** and **change with time**

Medical claims are **knowledge intensive**, and require careful reasoning of **information from many sources**

Question & answering; allows an **intuitive interaction** with the **medical knowledge** by utilizing **generative AIs**, saves **time** for the user and can easy the **understanding** of complex explanations

T-MQA: Research Questions

RQ1: What is the best performing **approach** for **medical question answering** and do these approaches **generalize** well over **diverse (or unseen) datasets**?

RQ2: How can we accurately generate **answers** to medical questions using **retrieved medical evidence (or knowledge)** using **LLMs** and the **RAG** method (Retrieval-augmented generation)?

RQ3: Can we **generate medically accurate explanations** in a Q&A format for users to understand medical information easier?

❖ T-MQA: Research Question 1 - Approach

RQ1: What is the best performing **approach** for **medical question answering** and do these approaches **generalize** well over **diverse (or unseen) datasets**?

- Among 15 medical datasets dating 2017 and after:
- categorization based on MQA-Survey of [Jin et al.]

	Clinical	**Examination**	**Scientific**	**Consumer**
Question	Does patient have abnormal BMI?	Antibiotics can be used to treat _?_. (e.g. MCQ: A/B/C)	Helicases are motor proteins that unwind _?_.	Can asthma be cured?
Answer	BMI: 31.2, Yes	C. Bacterial infections	nucleic acid	Asthma is chronic. It can be treated, but not cured.
Dataset	emrQA, CliCR	HEAD-QA	SciFact, **BioASQ**	HealthVer, PubHealth

T-MQA: Research Question 1 - Approach

RQ1: What is the best performing **approach** for **medical question answering** and do these approaches **generalize** well over **diverse (or unseen) datasets**?

→ BioASQ based Question Categories
→ Answer length

Average length
for most datasets

challenging for different
categories to unify e.g.
consumer/research

Chance of
hallucination
increases

Yes/No	**List**	**Factoid**	**Summary**
Does patient have abnormal BMI?	Antibiotics can be used to treat _?_. (e.g. MCQ)	Helicases are motor proteins that unwind _?_.	Can asthma be cured?
BMI: 31.2, Yes	C. Bacterial infections	nucleic acid	Asthma is chronic. It can be treated, but not cured.

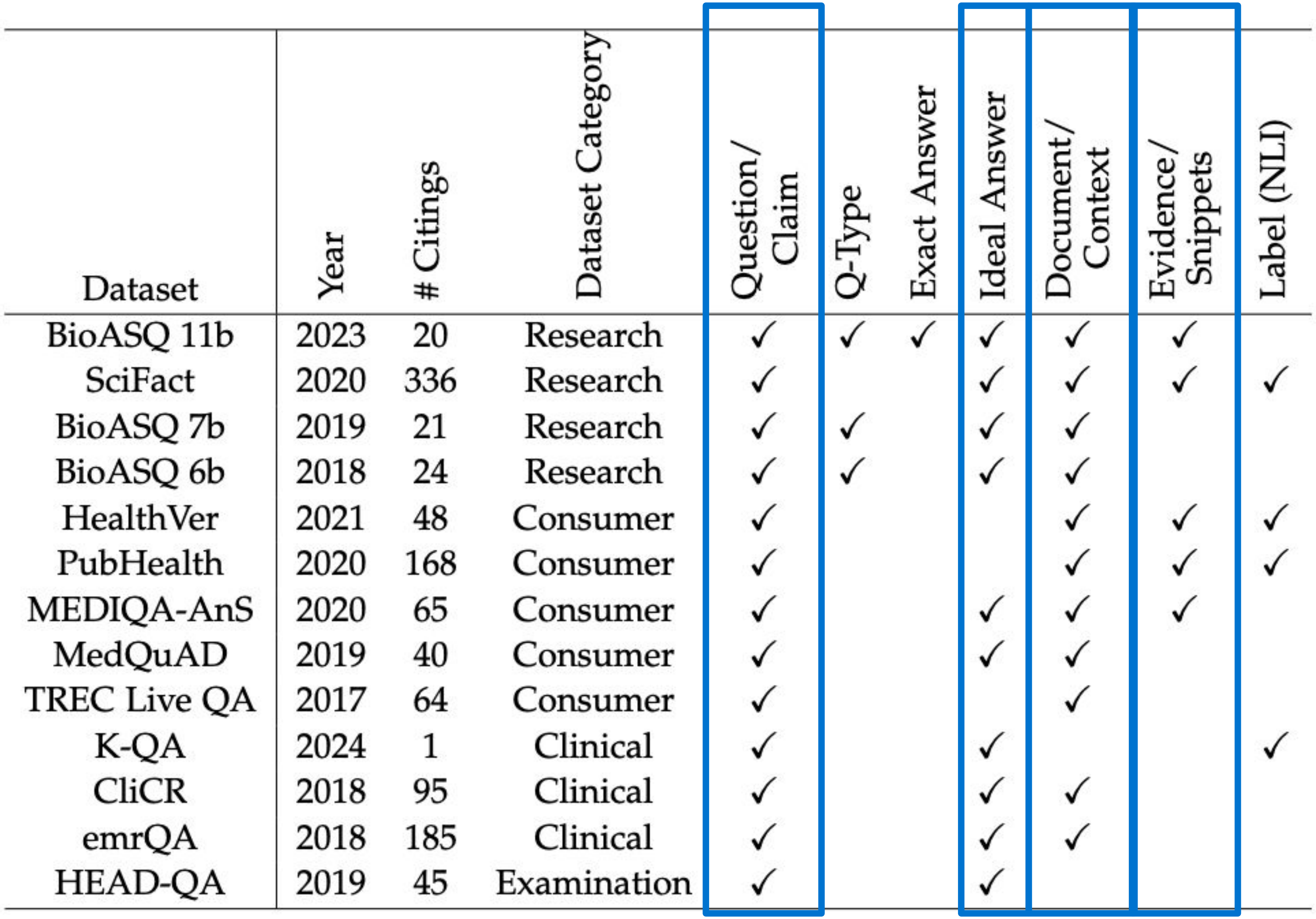
T-MQA: Research Question 1 - Approach

Dataset	Year	# Citings	Dataset Category	Question/ Claim	Q-Type	Exact Answer	Ideal Answer	Document/ Context	Evidence/ Snippets	Label (NLI)
BioASQ 11b	2023	20	Research	✓	✓	✓	✓	✓	✓	
SciFact	2020	336	Research	✓			✓	✓	✓	✓
BioASQ 7b	2019	21	Research	✓	✓		✓	✓		
BioASQ 6b	2018	24	Research	✓	✓		✓	✓		
HealthVer	2021	48	Consumer	✓				✓	✓	✓
PubHealth	2020	168	Consumer	✓				✓	✓	✓
MEDIQA-AnS	2020	65	Consumer	✓			✓	✓	✓	
MedQuAD	2019	40	Consumer	✓			✓	✓		
TREC Live QA	2017	64	Consumer	✓				✓		
K-QA	2024	1	Clinical	✓			✓			✓
CliCR	2018	95	Clinical	✓			✓	✓		
emrQA	2018	185	Clinical	✓			✓	✓		
HEAD-QA	2019	45	Examination	✓			✓			

Claims can be converted to questions using the labels

Q+A ≡ C+L

❖ T-MQA: Research Question 1 - Approach



Dataset	Year	# Citings	Dataset Category	Question/ Claim	Q-Type	Exact Answer	Ideal Answer	Document/ Context	Evidence/ Snippets	Label (NLI)
BioASQ 11b	2023	20	Research	✓	✓	✓	✓	✓	✓	
SciFact	2020	336	Research	✓			✓	✓	✓	✓
BioASQ 7b	2019	21	Research	✓	✓		✓	✓		
BioASQ 6b	2018	24	Research	✓	✓		✓	✓		
HealthVer	2021	48	Consumer	✓				✓	✓	✓
PubHealth	2020	168	Consumer	✓				✓	✓	✓
MEDIQA-AnS	2020	65	Consumer	✓			✓	✓	✓	
MedQuAD	2019	40	Consumer	✓			✓	✓		
TREC Live QA	2017	64	Consumer	✓				✓		
K-QA	2024	1	Clinical	✓			✓			✓
CliCR	2018	95	Clinical	✓			✓	✓		
emrQA	2018	185	Clinical	✓			✓	✓		
HEAD-QA	2019	45	Examination	✓			✓			

Claims will be converted to questions using the labels

Q+A ≡ C+L

T-MQA: Research Question 1 - Approach

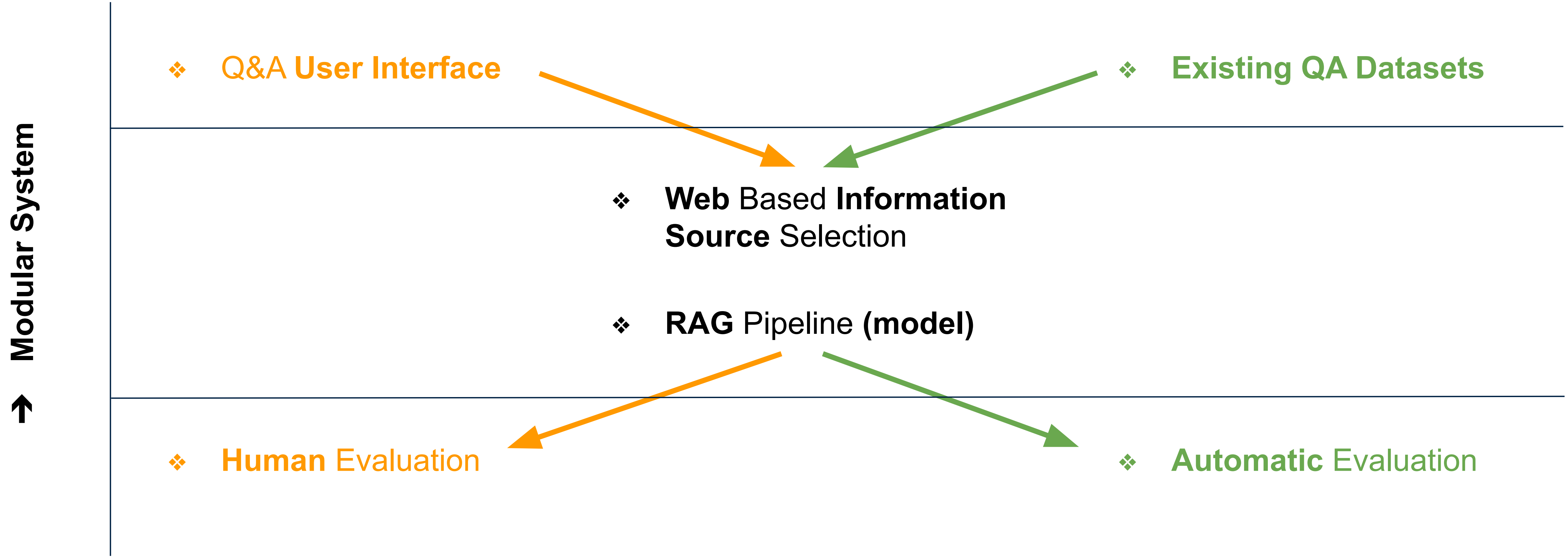
RQ1: What is the best performing **approach** for **medical question answering** and do these approaches **generalize** well over **diverse (or unseen) datasets**?

Example Q: **What disease is tinea?**

Golden Answer (BioASQ 11b)	**ChatGPT (Gpt-4-turbo-preview)**	**HuggingChat (c4ai-command-r-plus)+ Web(SerpAPI:Google)**	**Perplexity.ai** (Llama 3)	**GPT3.5+HyDE +RAG(sofar only wiki)**
Tinea is a superficial fungal infections of the skin.	Tinea is a fungal infection of the skin, also known as ringworm. It can affect various body parts and manifests as a red, itchy, circular rash with clearer skin in the middle.	Tinea is a fungal infection of the skin. It is also known as ringworm	Tinea is a contagious fungal skin infection that can affect various parts of the body, including the feet, groin, scalp, and nails. It is also known as ringworm, …	Tinea is a type of fungal infection. It can affect various parts of the body, including the skin, hair, and nails. Tinea is also known as ringworm due to the characteristic ring-like shape of the rash that appears on the affected area.

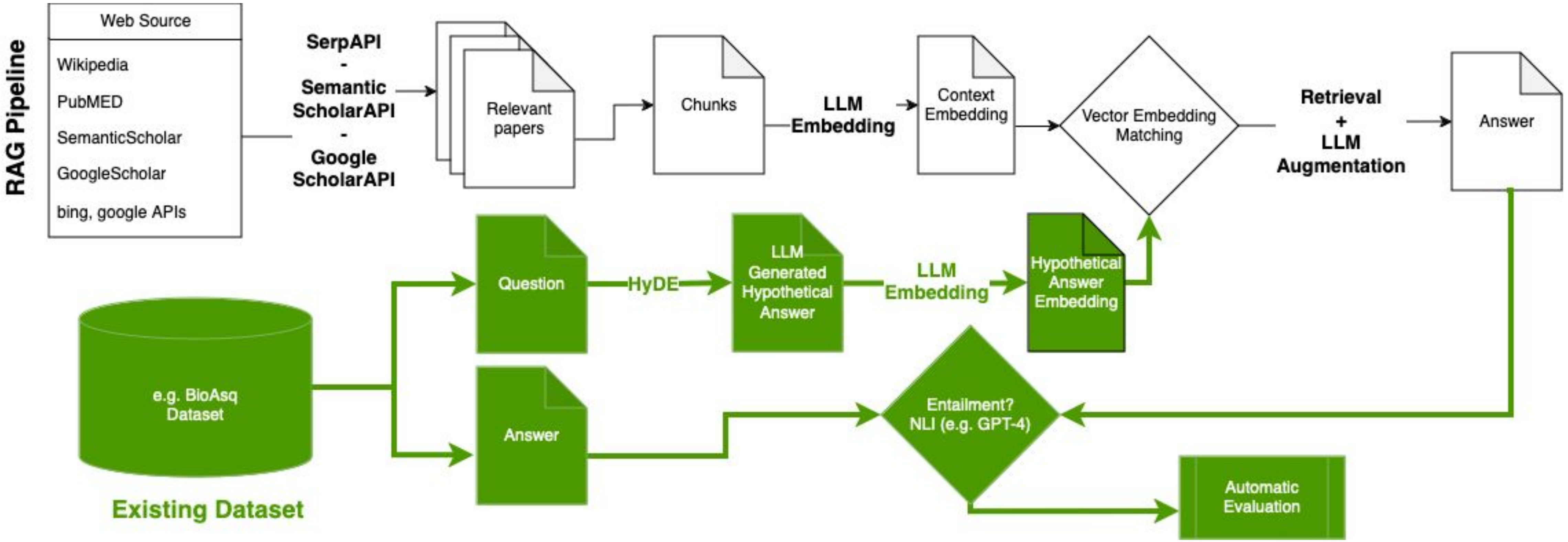
T-MQA: Research Question 2 - System

RQ2: How can we accurately generate **answers** to medical questions using **retrieved medical evidence (or knowledge)** using **LLMs** and the **RAG** method (Retrieval-augmented generation)?

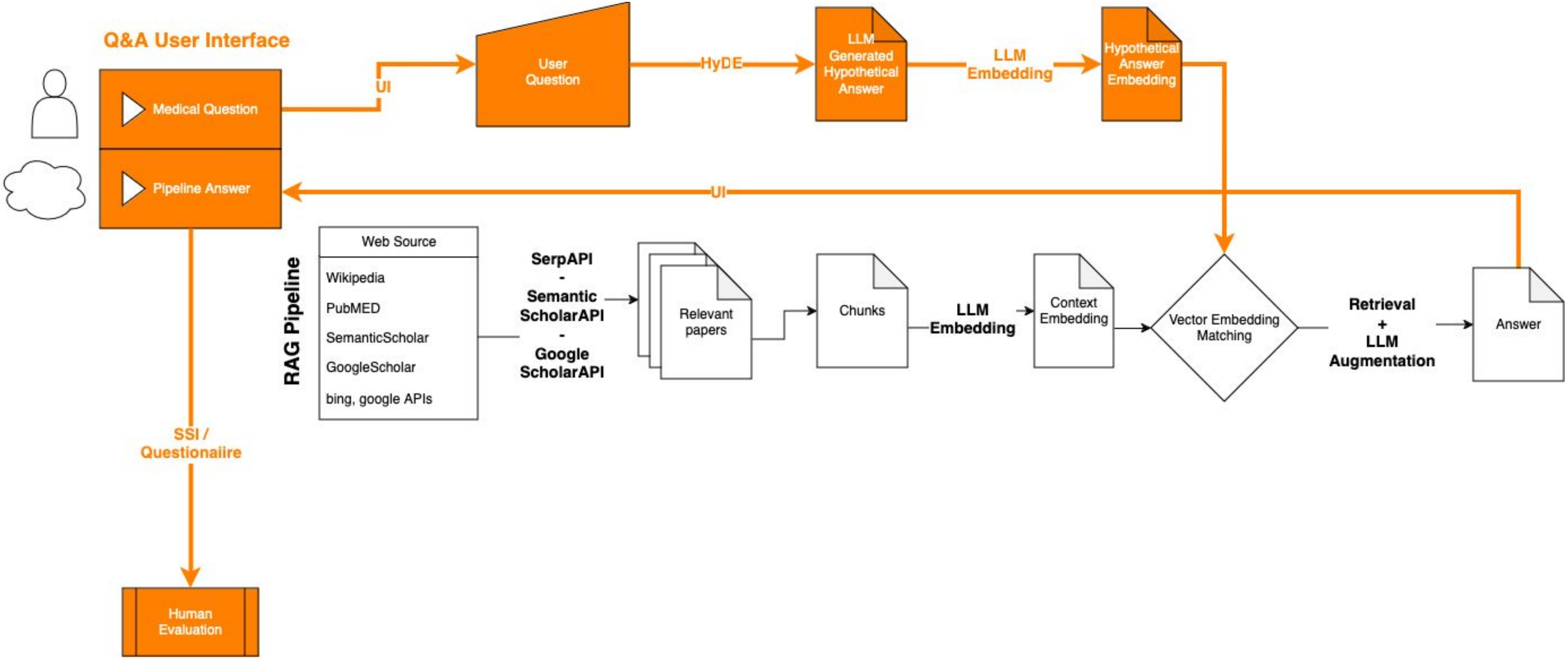


- Q&A **User Interface**
- **Existing QA Datasets**
- **Web Based Information Source** Selection
- **RAG** Pipeline **(model)**
- **Human** Evaluation
- **Automatic** Evaluation

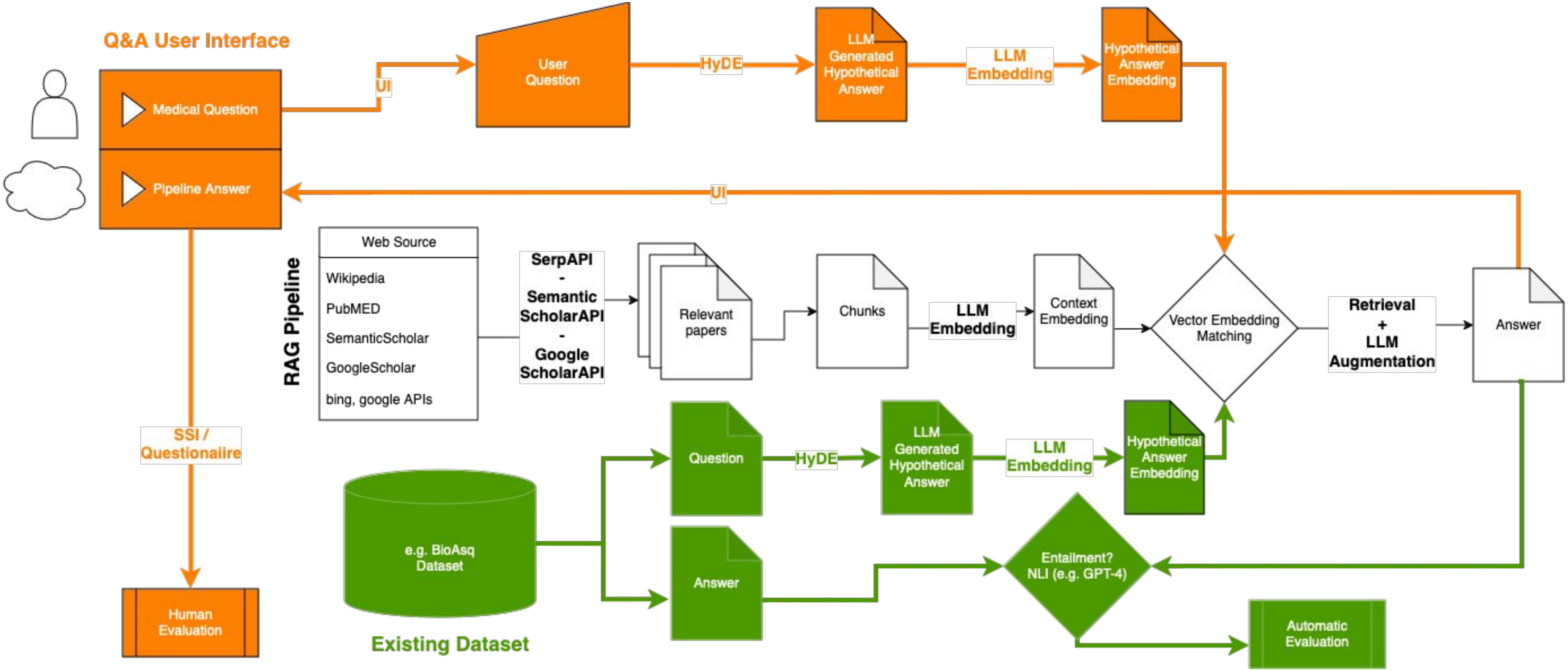
T-MQA: Research Question 2 - System



T-MQA: Research Question 2 - System



T-MQA: Research Question 2 - System



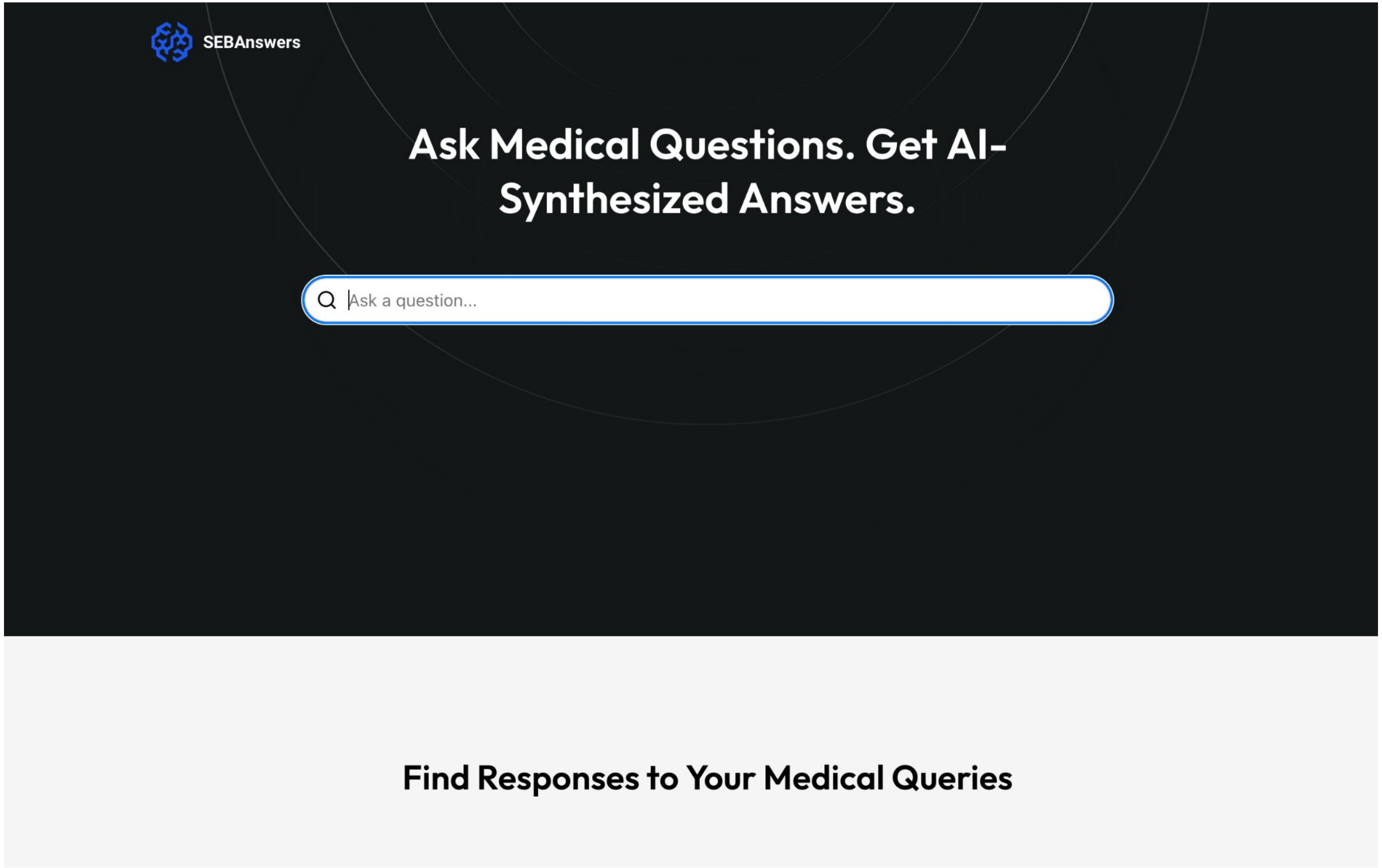
T-MQA: Research Question 2 - System

RQ2: How can we accurately generate **answers** to medical questions using **retrieved medical evidence (or knowledge)** using **LLMs** and the **RAG** method (Retrieval-augmented generation)?

- **Automatic** Evaluation with **existing Datasets**
  - MRR
  - F1
  - …
  - NLI (Entailment)

- Human Evaluation
  - Questionaire/Interview
  - Manual

T-MQA: Research Question 3 - Users

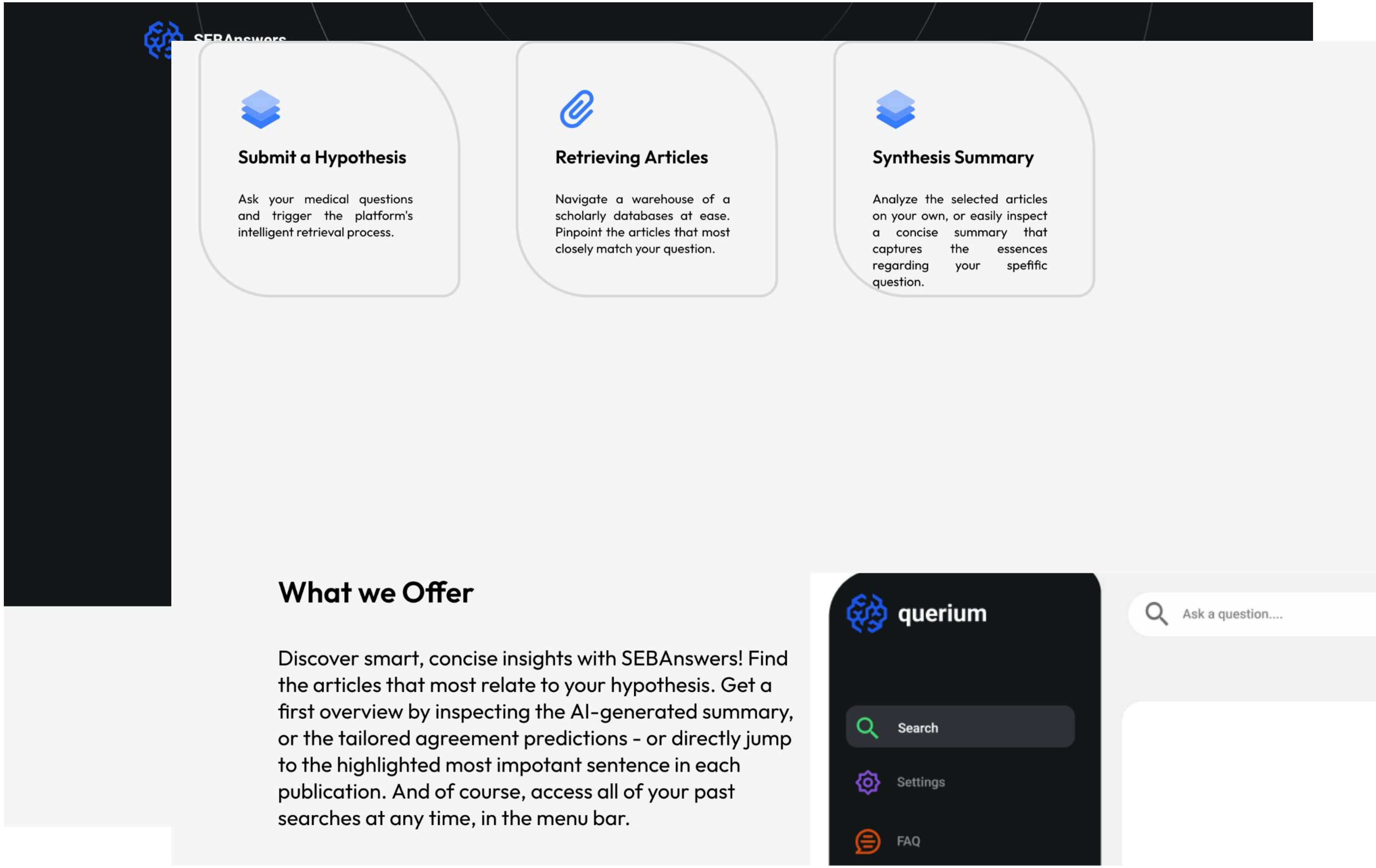


SEBAnswers

Ask Medical Questions. Get AI-Synthesized Answers.

Find Responses to Your Medical Queries

T-MQA: Research Question 3 - Users



Submit a Hypothesis

Ask your medical questions and trigger the platform's intelligent retrieval process.

Retrieving Articles

Navigate a warehouse of a scholarly databases at ease. Pinpoint the articles that most closely match your question.

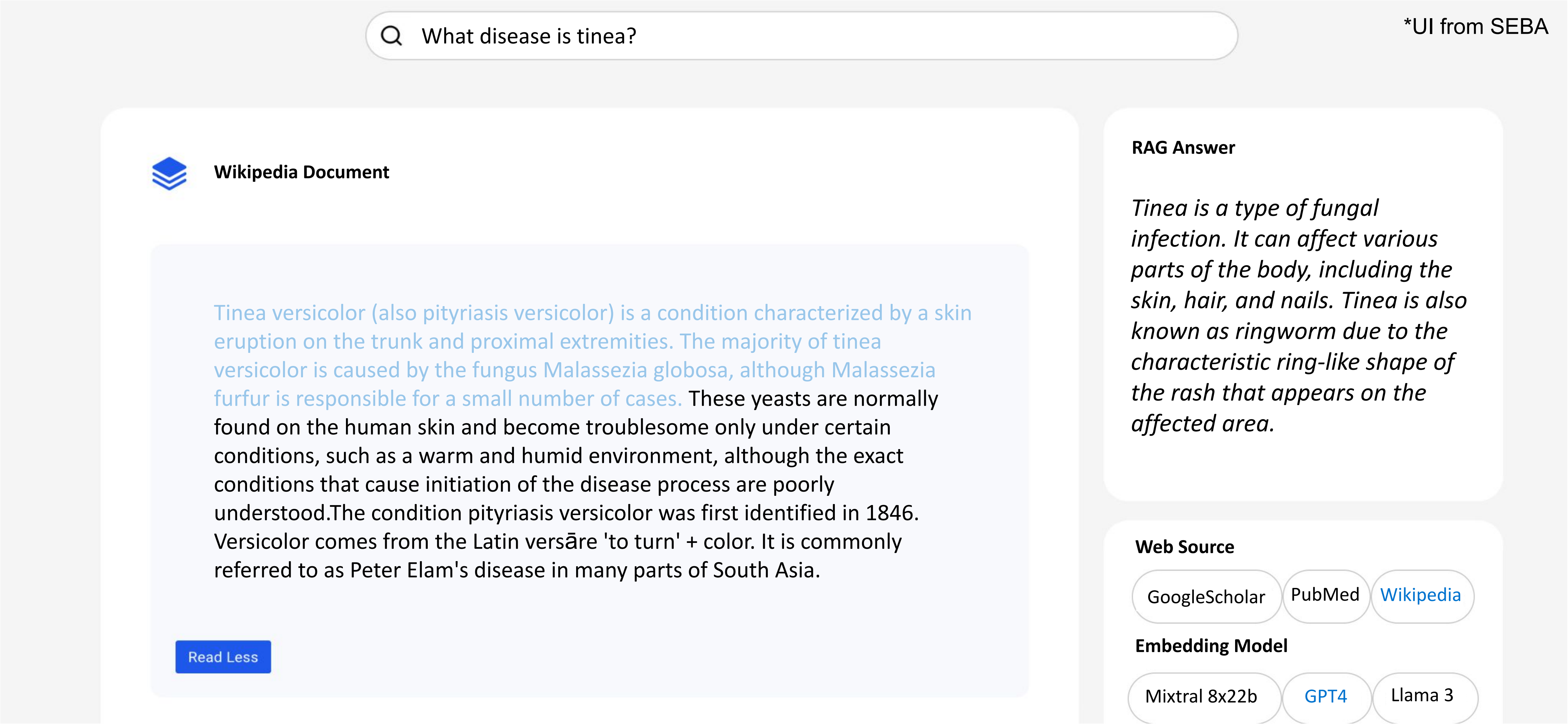
Synthesis Summary

Analyze the selected articles on your own, or easily inspect a concise summary that captures the essences regarding your spefific question.

What we Offer

Discover smart, concise insights with SEBAnswers! Find the articles that most relate to your hypothesis. Get a first overview by inspecting the AI-generated summary, or the tailored agreement predictions - or directly jump to the highlighted most impotant sentence in each publication. And of course, access all of your past searches at any time, in the menu bar.

T-MQA: Research Question 3 - Users



*UI from SEBA

What disease is tinea?

Wikipedia Document

Tinea versicolor (also pityriasis versicolor) is a condition characterized by a skin eruption on the trunk and proximal extremities. The majority of tinea versicolor is caused by the fungus Malassezia globosa, although Malassezia furfur is responsible for a small number of cases. These yeasts are normally found on the human skin and become troublesome only under certain conditions, such as a warm and humid environment, although the exact conditions that cause initiation of the disease process are poorly understood.The condition pityriasis versicolor was first identified in 1846. Versicolor comes from the Latin versāre 'to turn' + color. It is commonly referred to as Peter Elam's disease in many parts of South Asia.

RAG Answer

Tinea is a type of fungal infection. It can affect various parts of the body, including the skin, hair, and nails. Tinea is also known as ringworm due to the characteristic ring-like shape of the rash that appears on the affected area.

Web Source

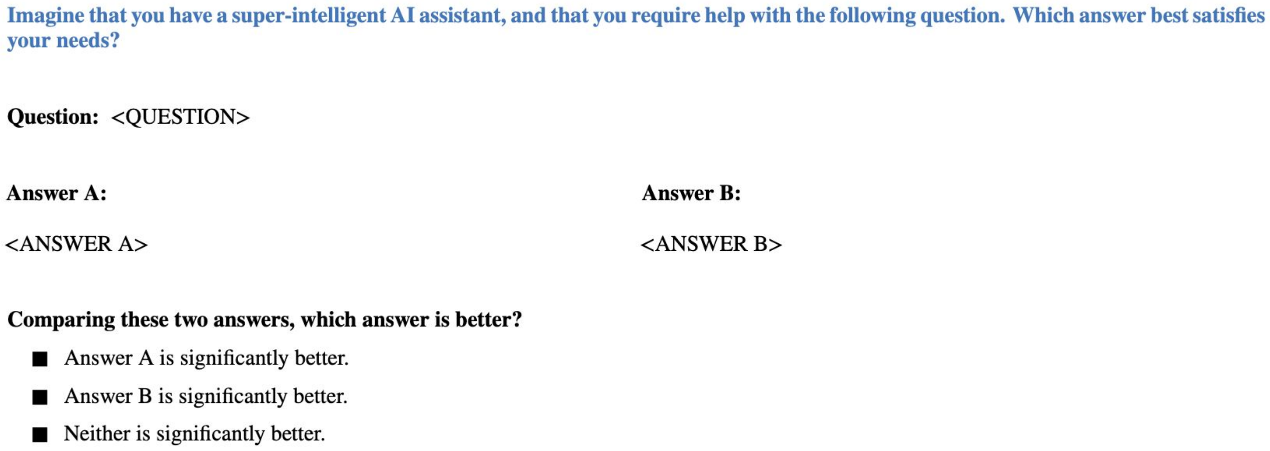
Embedding Model

T-MQA: Research Question 3 - Users

RQ3: Can we **generate medically accurate explanations** in a Q&A format for users to understand medical information easier?

➔ Questionnaire: Blind human evaluation of differences of the following to examine impact on trustworthiness:
- dataset/web sources
- embedding models
- generated answers to golden answer

Human annotation interface



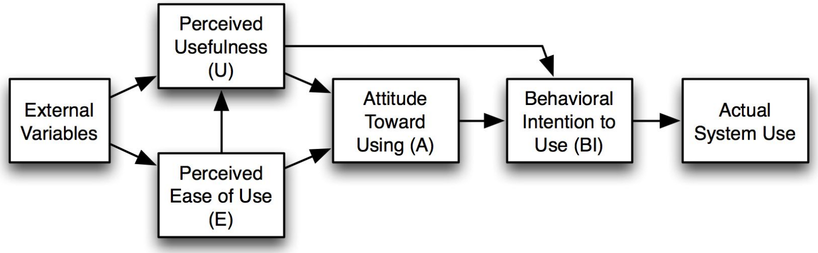
Imagine that you have a super-intelligent AI assistant, and that you require help with the following question. Which answer best satisfies your needs?

Question: <QUESTION>

Answer A:	**Answer B:**
<ANSWER A>	<ANSWER B>

Comparing these two answers, which answer is better?
- ■ Answer A is significantly better.
- ■ Answer B is significantly better.
- ■ Neither is significantly better.

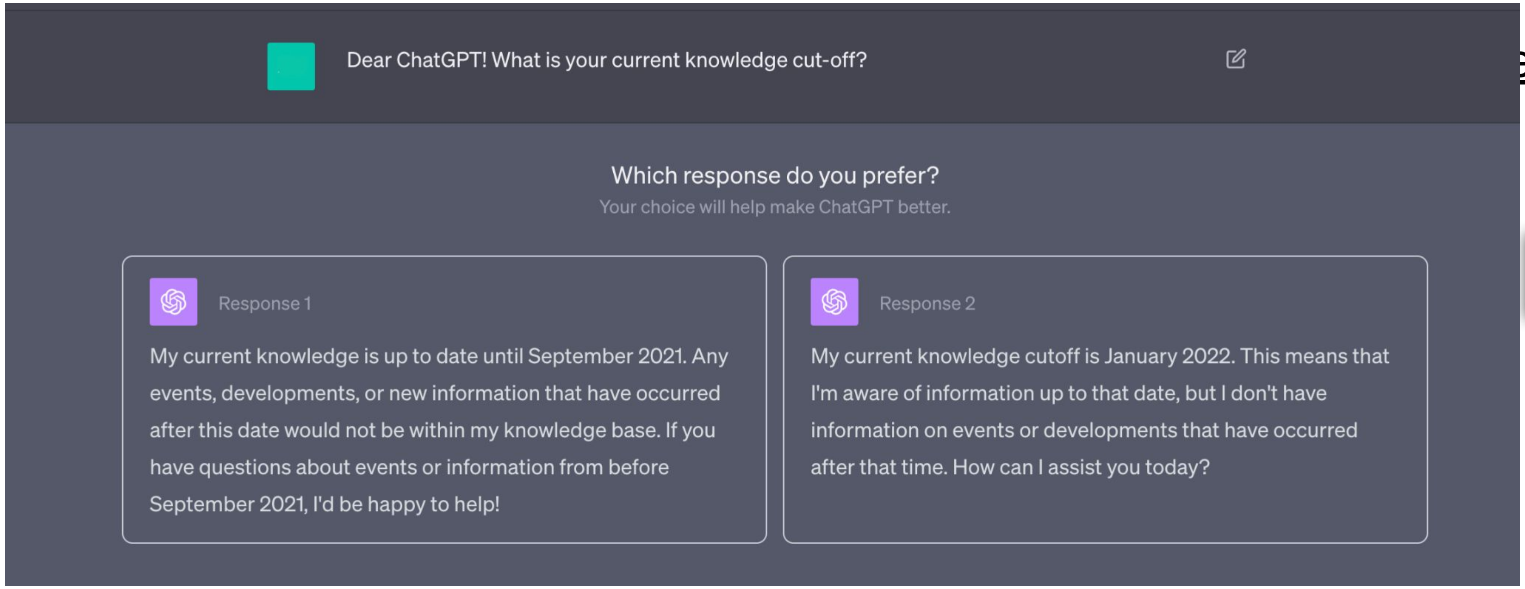
Technology acceptance model (optional):


External Variables
Perceived Usefulness (U)
Perceived Ease of Use (E)
Attitude Toward Using (A)
Behavioral Intention to Use (BI)
Actual System Use

❖ T-MQA: Research Question 3 - Users

RQ3: Can we **generate medically accurate explanations** in a Q&A format for users to understand medical information easier?

➔ Questionnaire: Blind human evaluation of differences of the following to examine impact on trustworthiness:
- dataset/web sources
- embedding models
- generated answers to golden answer



Dear ChatGPT! What is your current knowledge cut-off?

Which response do you prefer?

Your choice will help make ChatGPT better.

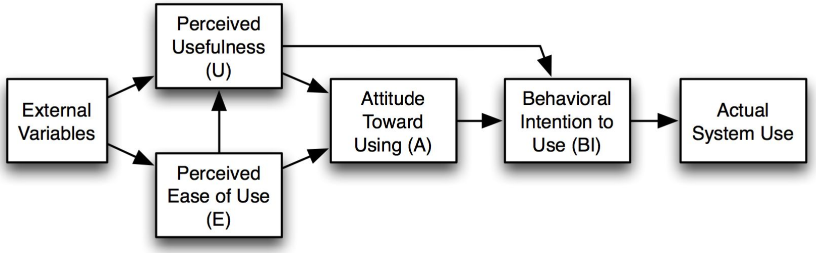
Response 1

My current knowledge is up to date until September 2021. Any events, developments, or new information that have occurred after this date would not be within my knowledge base. If you have questions about events or information from before September 2021, I'd be happy to help!

Response 2

My current knowledge cutoff is January 2022. This means that I'm aware of information up to that date, but I don't have information on events or developments that have occurred after that time. How can I assist you today?

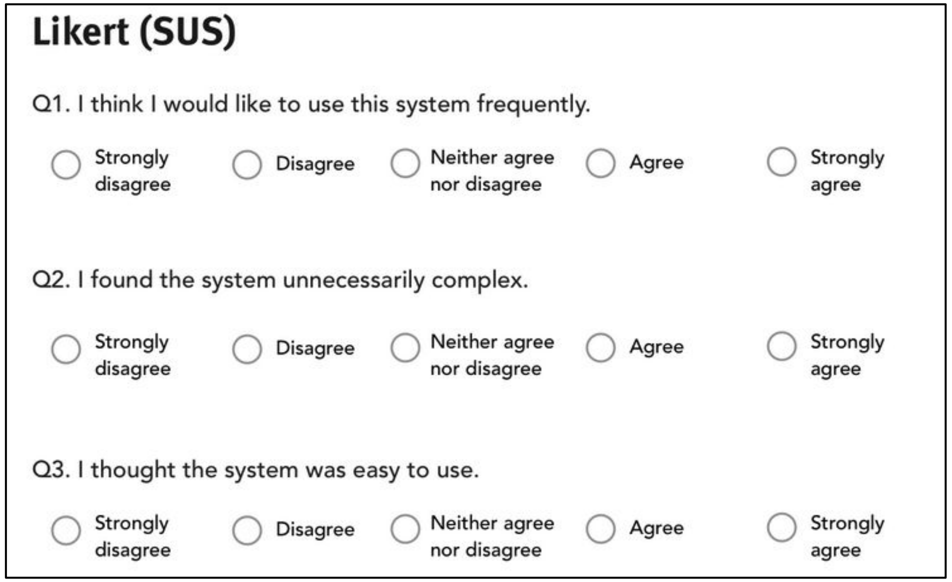
echnology acceptance model (optional):


External Variables
Perceived Usefulness (U)
Perceived Ease of Use (E)
Attitude Toward Using (A)
Behavioral Intention to Use (BI)
Actual System Use

T-MQA: Research Question 3 - Users

RQ3: Can we **generate medically accurate explanations** in a Q&A format for users to understand medical information easier?

- **Technical Human Evaluation**
  - G-EVAL (Spearman (ρ) and Kendall-Tau (τ))
    - Naturalness
    - Coherence
    - Engagingness
    - Groundedness
    - AVG



Likert (SUS)

Q1. I think I would like to use this system frequently.

○ Strongly disagree ○ Disagree ○ Neither agree nor disagree ○ Agree ○ Strongly agree

Q2. I found the system unnecessarily complex.

○ Strongly disagree ○ Disagree ○ Neither agree nor disagree ○ Agree ○ Strongly agree

Q3. I thought the system was easy to use.

○ Strongly disagree ○ Disagree ○ Neither agree nor disagree ○ Agree ○ Strongly agree

- **Preferential Human Evaluation**
  - Age, Nationality, Occupation,...
  - OPTION: Web source
    - PubMed, Scholar, Wikipedia
  - OPTION: AI model provider
    - Meta, OpenAI,
  - Control Group for OPTIONs? (e.g. best model with all sources)

❖ T-MQA: Research Question 3 - Users

RQ3: Can we **generate medically accurate explanations** in a Q&A format for users to understand medical information easier?

➔ Possible **UI extensions** (nice to have)

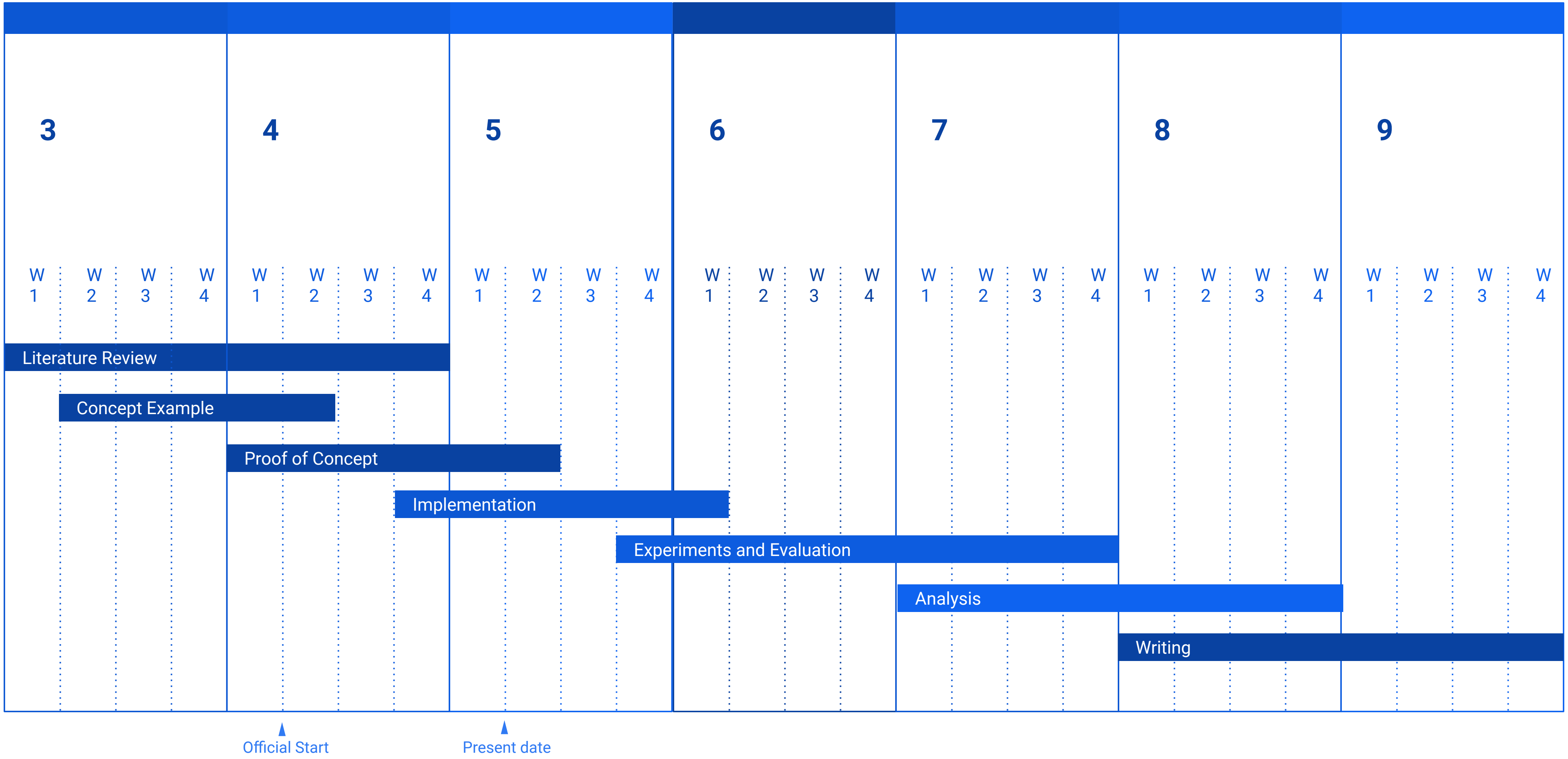
Medical Report Upload

German - English **Translation**

DeepL

Text-to-Speech ←→ Speech-to-Text

Timeline


3
4
5
6
7
8
9
W 1
W 2
W 3
W 4
Literature Review
Concept Example
Proof of Concept
Implementation
Experiments and Evaluation
Analysis
Writing
Official Start
Present date

B.Sc.

Ragip Volkan Tatlikazan

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.——
volkan.tatlikazan@.tum.de
wwwmatthes.in.tum.de