Developing Systems for Trustworthy Medical Question Answering

Ragip Volkan Tatlikazan

2024.10.21

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

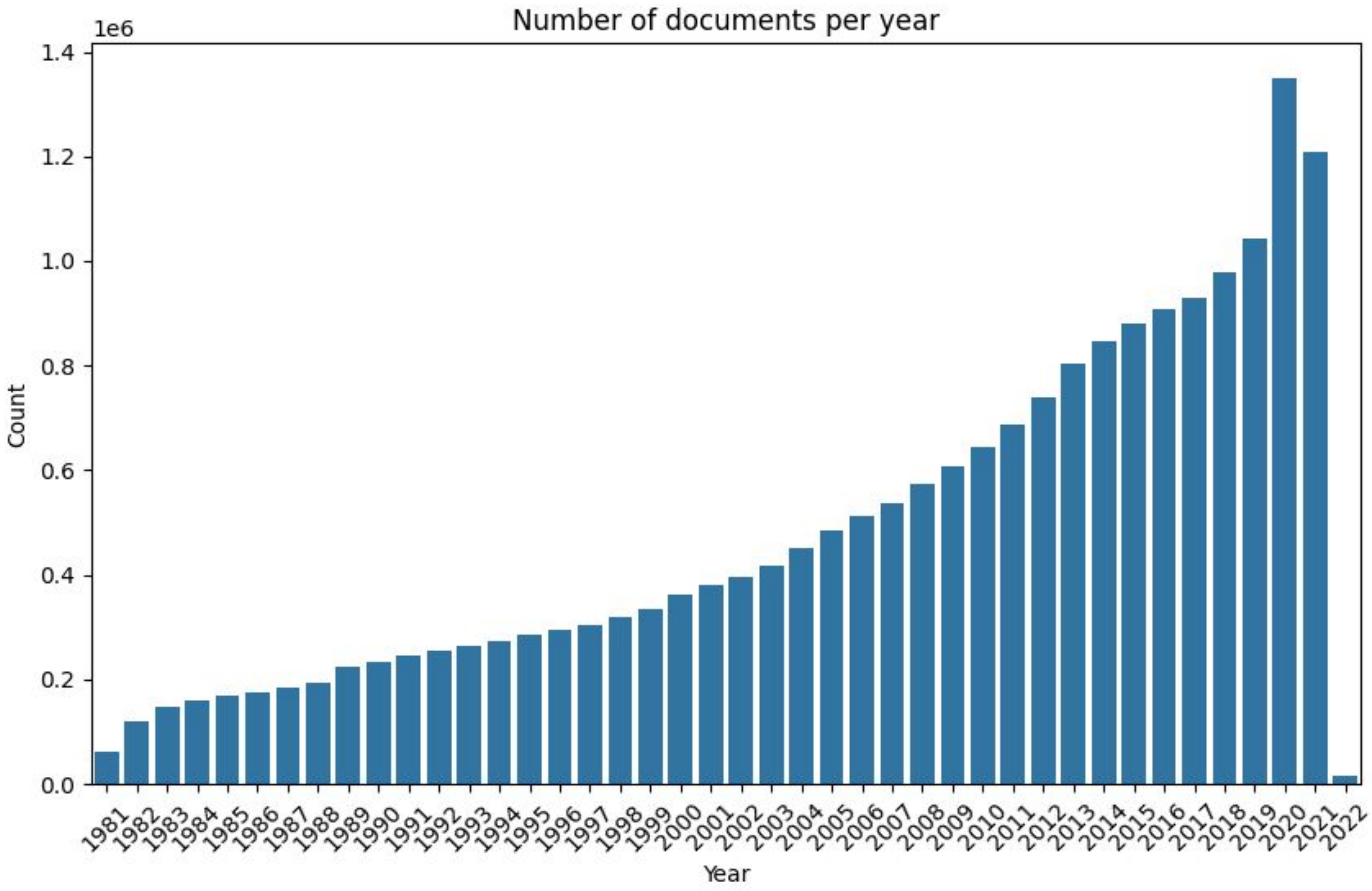
- Developing Systems for Trustworthy Medical Question Answering (T-MQA)
  - Motivation
  - Research Questions
  - System & Experiments
  - Human Evaluation
  - Key Takeaways
  - Future Work

❖ T-MQA: Motivation

❖ T-MQA: Motivation

Medical-QA:

- Constantly advancing medical knowledge
  - invalidates older info
- Redacted heavily cited papers
  - Sometimes not even updated in some academic publisher websites
- **-> Need for a system**



*Nov 2022: ChatGpt release

❖ T-MQA: Motivation

Trustworthy -> Private + Traceable:

- Data breaches of big tech companies
  - raises concern for personal medical data

- Medical information websites/blogs/social media without sources

- **Solution -> LOCAL system + manual knowledge updates**

❖ T-MQA: Research Questions

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: Approach

	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	k-QA	-	**BioASQ**	HealthFC, AKI-Gen

- Dataset choices:
  - Examination discarded as **mostly MCQ**
  - BioASQ (experiments)
    - reliable and comparable, many features for future work
  - k-QA
    - recent, **high performance results**, rigorous expert answers (no gen.)
  - HealthFC
    - from SEBIS, easier understanding - integration
  - Alpha KI Gen
    - from SEBIS, **LLM generated Dataset, with expert checks**

T-MQA: Approach

- Framework choices:
  - Ollama vs Pytorch (hf, transformers, ..)
    - C++ vs Python
    - less RAM requirements

- Batch Processes (Embedding/Inference) -> currently Sequentially
  - **M3 Max**: no Metall GPU support
  - **Sebis - Nvidia V100**: not enough RAM

- Vector Store:
  - FAISS (open-source)
  - VectorDBs (perpetual updates)
    - Weaviate

- Web sources
  - PubMED (20M abstracts)
  - Wikipedia (6M pages)

Speed Test BM25

Google Colab has proven to have unreliable connection.

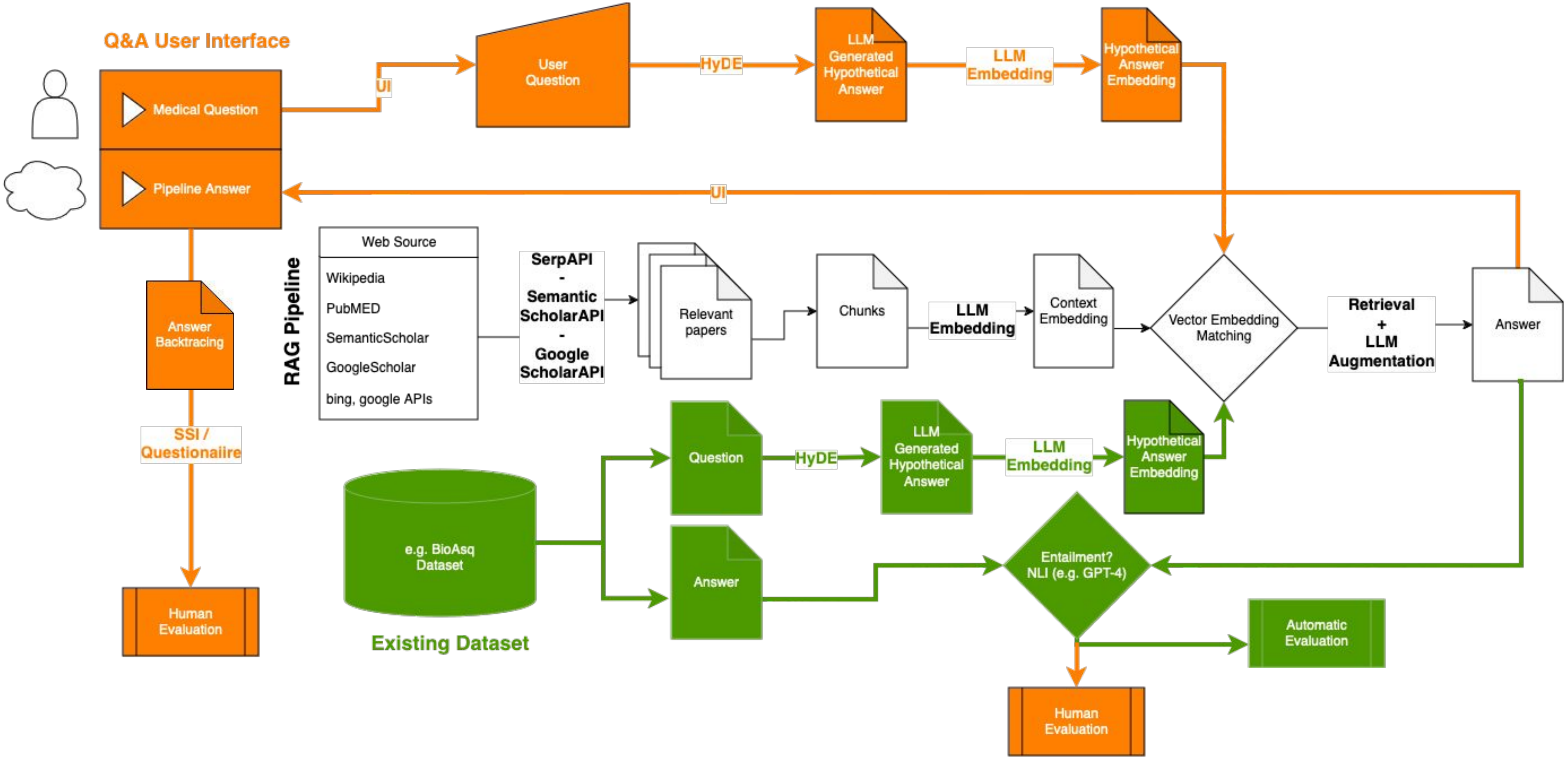
Device	Num. Docs	Num. QA Pairs	Duration
M3	20,000,000	10	0:40:56
M3	10,000,000	10	0:19:08
M3	1,000,000	10	0:02:39
M3	100,000	10	0:01:02
M3	10,000	10	0:00:44
M3	1,000	10	0:00:44
M3	1,000	1,000	2:33:59

T-MQA: Approach

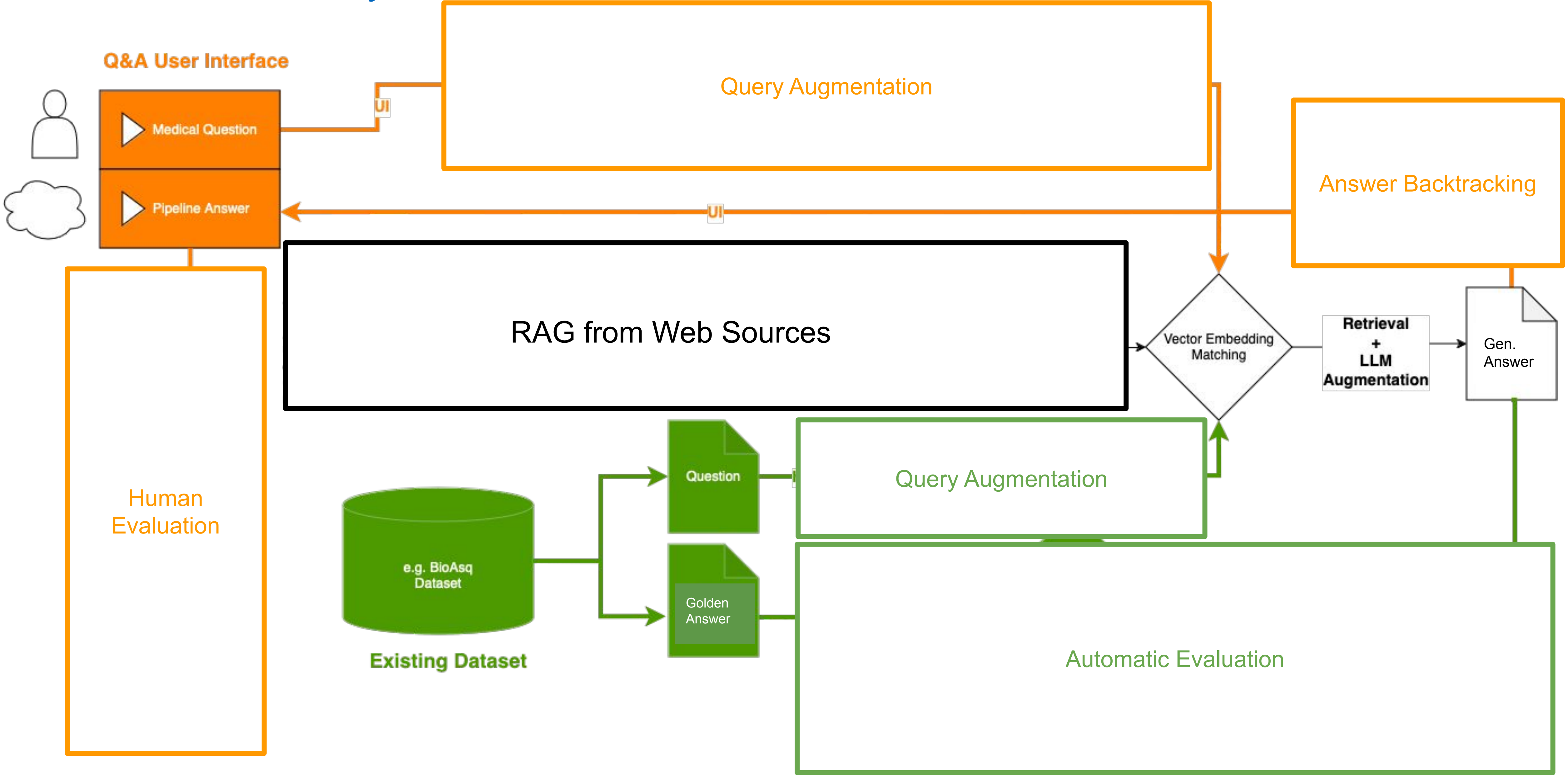
- Automatic Evaluation Metrics:
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
    - ROUGE compares an automatically produced summary or translation against a reference or set of reference (human-produced) summaries or translations
  - BART (Bidirectional and Auto-Regressive Transformers)
    - BARTScore uses pre-trained sequence-to-sequence models
      - can be applied in an unsupervised manner
- Human Evaluation
  - Questionnaire (152 respondant)
  - Manual Annotation (Supervisor & Student)
  - Short Interview (2 interviews)

- T-MQA: System & Experiments

T-MQA: RAG System



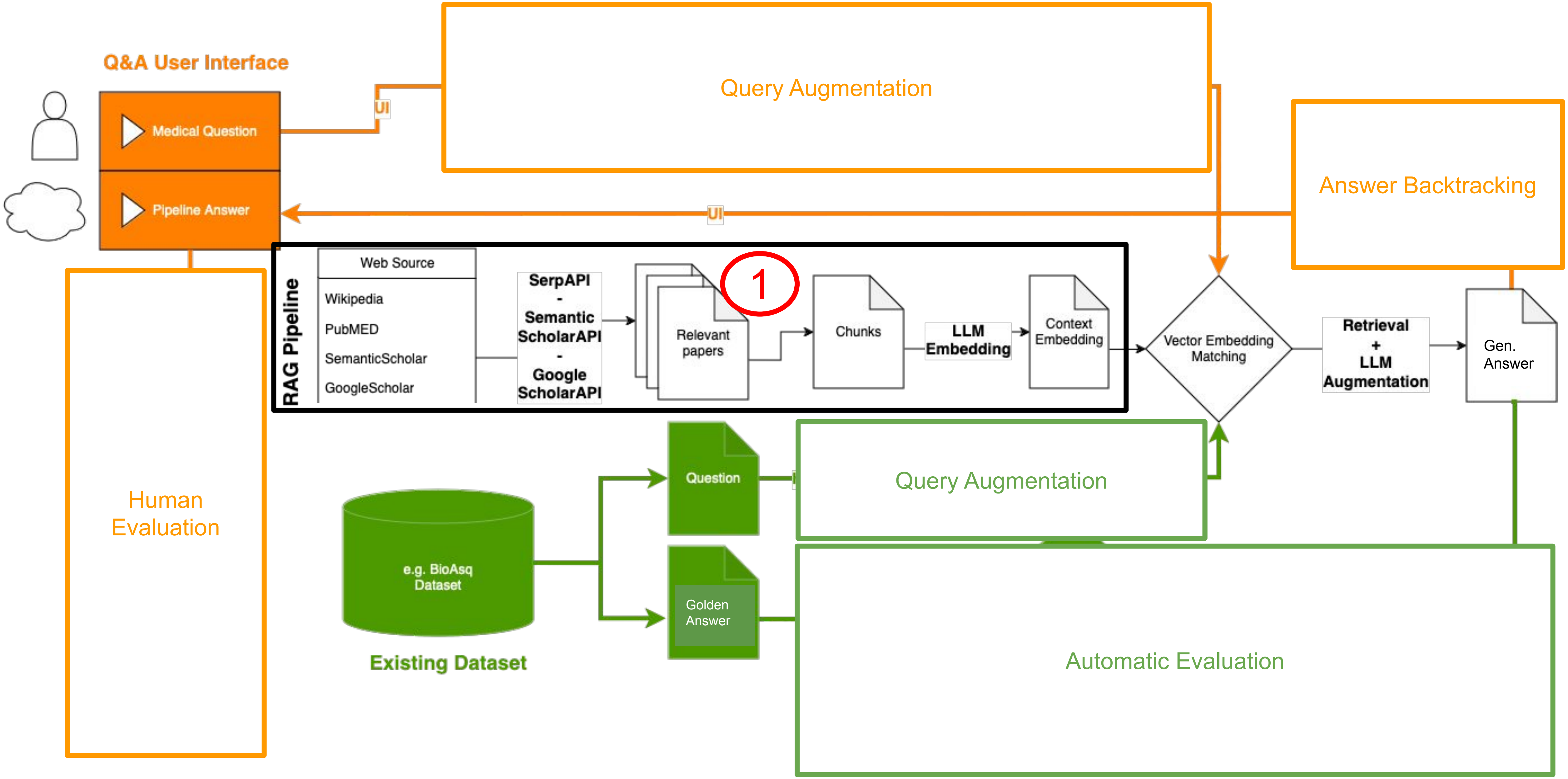
T-MQA: RAG System



T-MQA: Experiments

1. Number of **Retrieved Documents** Test
2. **Query Augmentation**: HyDE Test
3. LLM **Inference** Test
4. **Keyword + Semantic Embedding** Test
5. **Pubmed vs Wikipedia** Inference Test
6. **Keyword frequency** Test (BM25)
7. **Automatic vs Human Evaluation** Test

T-MQA: Number of Retrieved Documents



Number of Retrieved Documents Test

- M3
  - 20 Million PubMED Abstracts
  - 1000 BioASQ Questions

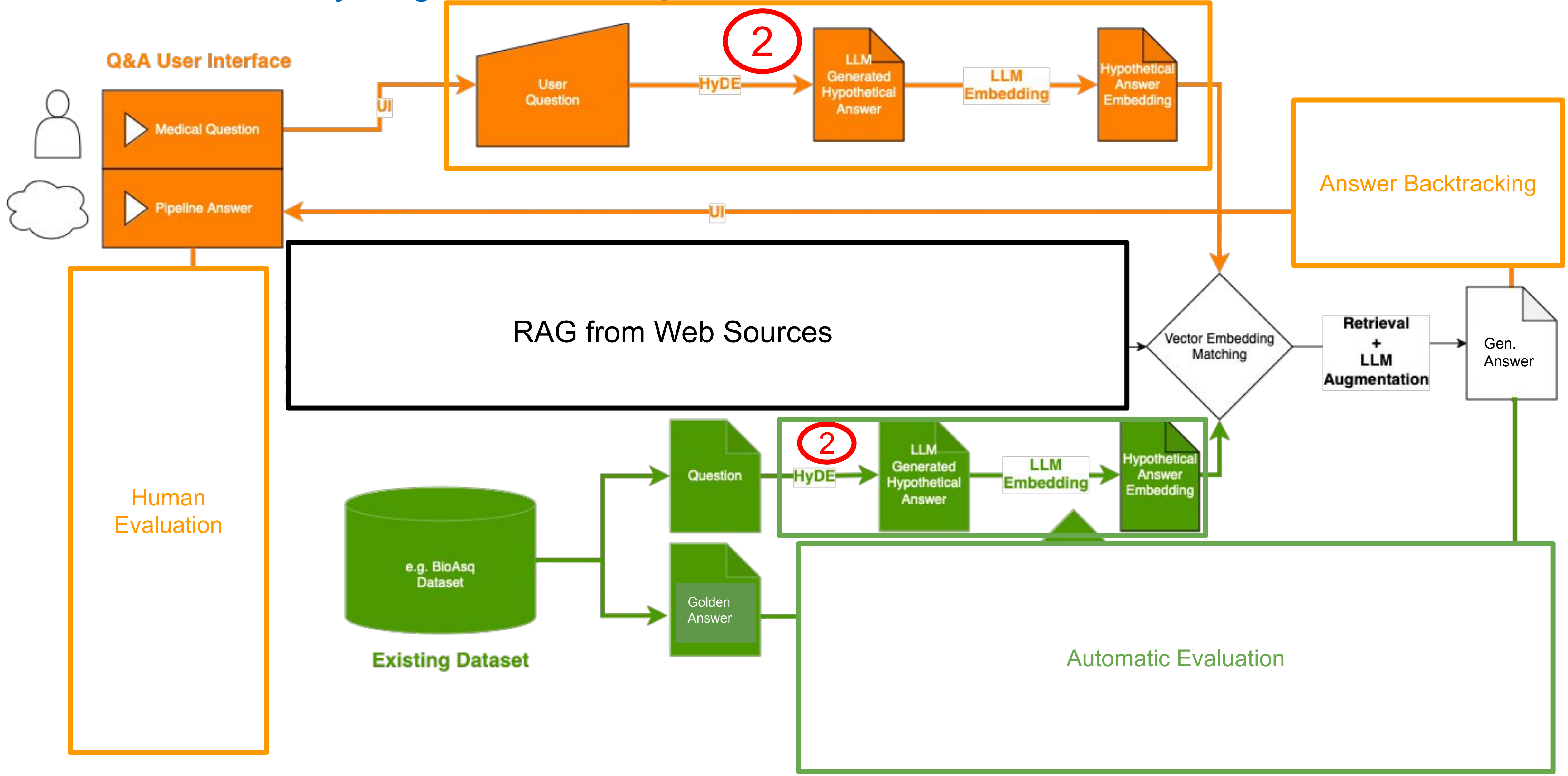
  [58] Test results support lost in the middle for increasing number of documents.

Number of retrieved Documents	1	3	5	7	9
rouge_scores.rouge1	25.17	27.43	**28.36**	27.30	25.16
rouge_scores.rouge2	8.54	10.55	**11.19**	9.95	7.78
rouge_scores.rougeL	18.06	20.12	**20.64**	19.54	17.54
rouge_scores.rougeLsum	18.12	20.28	**21.07**	19.99	17.83
bart_scores_avg	5.52	**6.38**	6.18	5.78	4.80

Table 6.1: Performance metrics across different numbers of retrieved documents. (in percentages)

- Odd number of documents
- Optimal is 3-5
- 1: not enough info
- 9: non relevant info
- We keep 5
  - to have most info with performance

T-MQA: Query Augmentation: HyDE Test



Hyde

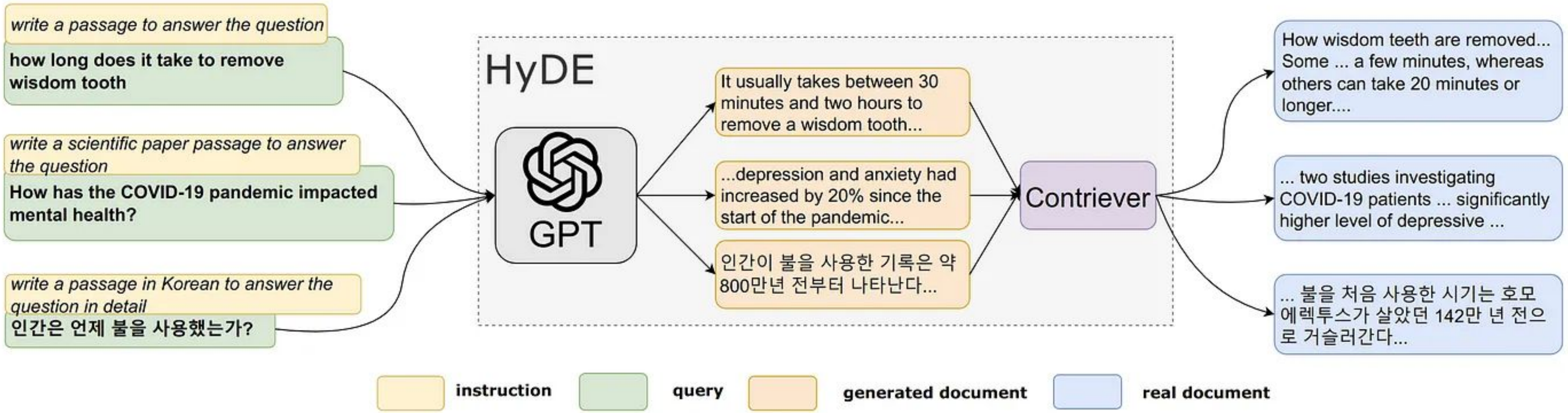


Figure 1: An illustration of the HyDE model. Documents snippets are shown. HyDE serves all types of queries without changing the underlying GPT-3 and Contriever/mContriever models.

HyDE Test

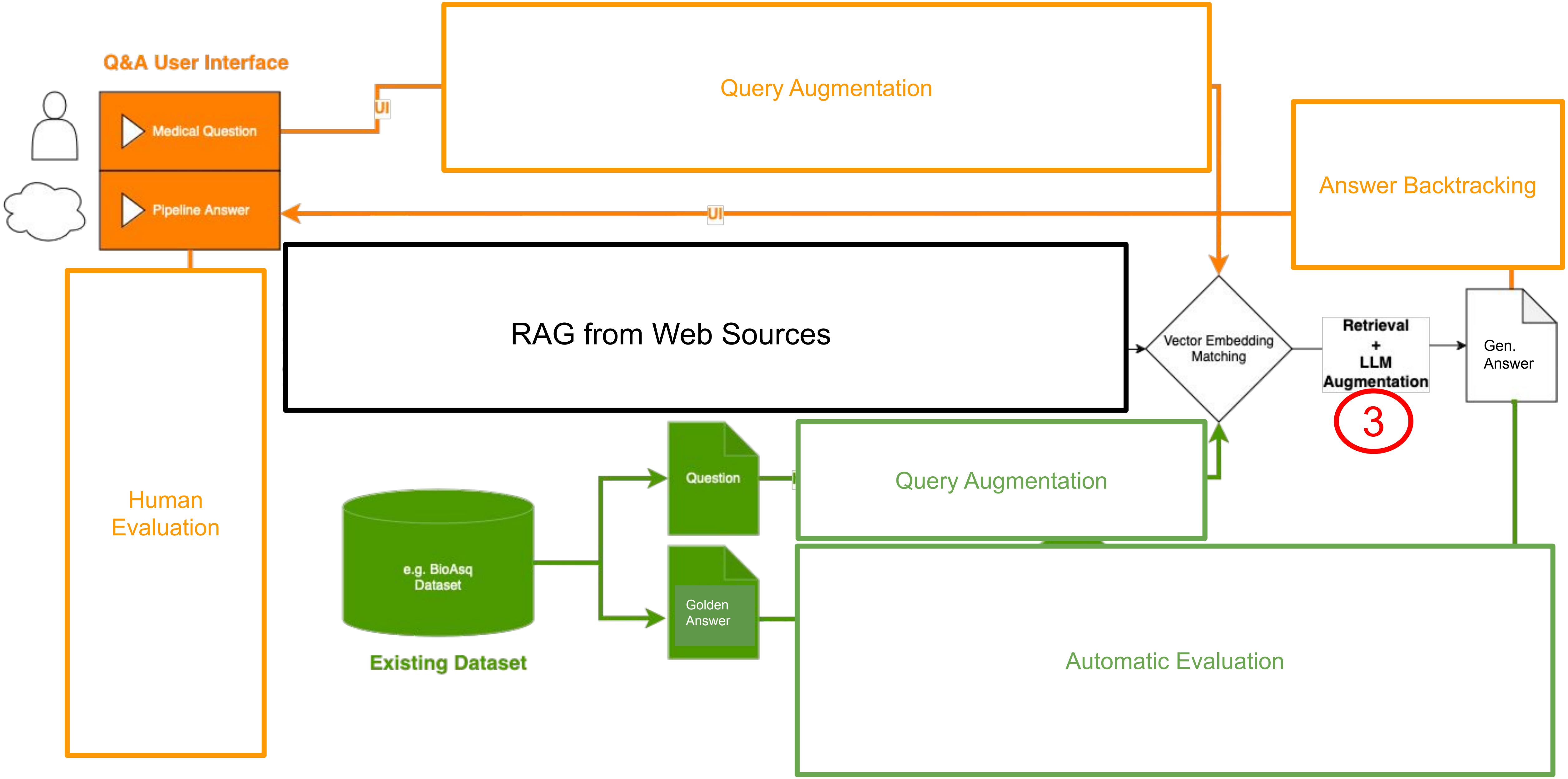
- M3
  - 20 Million PubMED Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (only BM25)
  - Inference LLM: Llama3:8b

apply_HyDE	FALSE	TRUE	FALSE	TRUE
embedding_model	nomic	nomic	mxbai	mxbai
bart_scores_avg	3.68	**4.22**	3.97	4.08
rouge_scores.rouge1	21.29	23.45	22.07	**23.52**
rouge_scores.rouge2	5.30	**6.70**	5.76	6.61
rouge_scores.rougeL	14.68	**16.48**	15.47	16.41
rouge_scores.rougeLsum	14.87	**16.76**	15.75	16.73

Table 6.4: Performance metrics across different embedding models with HyDE on/off. (in percentages)

- Tested on BERT embedding variants
- increase in performance regardless of model

T-MQA: LLM Inference Test



LLM Inference Test

- M3
  - Hyde: On
  - 20 Million PubMED Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (only BM25)

Metric	l3.1:405b	gpt4turbo	mixtral:8x7b	mistral	l3-chatqa:8b	l3.1:8b	l3:8b
bart_scores_avg	8.06	6.48	5.77	5.88	7.62	**8.12**	6.47
rouge1	29.63	**29.71**	28.24	28.76	19.98	27.84	28.45
rouge2	12.15	10.41	10.47	10.47	7.63	11.10	**11.32**
rougeL	**21.73**	20.26	19.76	20.11	16.04	20.64	21.05
rougeLsum	**21.72**	20.26	19.87	20.21	16.06	20.86	21.43

Table 6.5: Performance metrics across different LLM Inferences. (in percentages)

- Significant increase in model size doesnt effect the RAG system
- More recent models have better training, so they also perform better
- Open/closed source difference minimal

LLM Inference Test

- M3
  - Hyde: On
  - 20 Million PubMED Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (only BM25)

Metric	l3.1:405b	gpt4turbo	mixtral:8x7b	mistral	l3-chatqa:8b	l3.1:8b	l3:8b
bart_scores_avg	8.06	6.48	5.77	5.88	7.62	**8.12**	6.47
rouge1	29.63	**29.71**	28.24	28.76	19.98	27.84	28.45
rouge2	12.15	10.41	10.47	10.47	7.63	11.10	**11.32**
rougeL	**21.73**	20.26	19.76	20.11	16.04	20.64	21.05
rougeLsum	**21.72**	20.26	19.87	20.21	16.06	20.86	21.43

Table 6.5: Performance metrics across different LLM Inferences. (in percentages)

- Significant increase in model size doesnt effect the RAG system
- More recent models have better training, so they also perform better
- Open/closed source difference minimal

LLM Inference Test

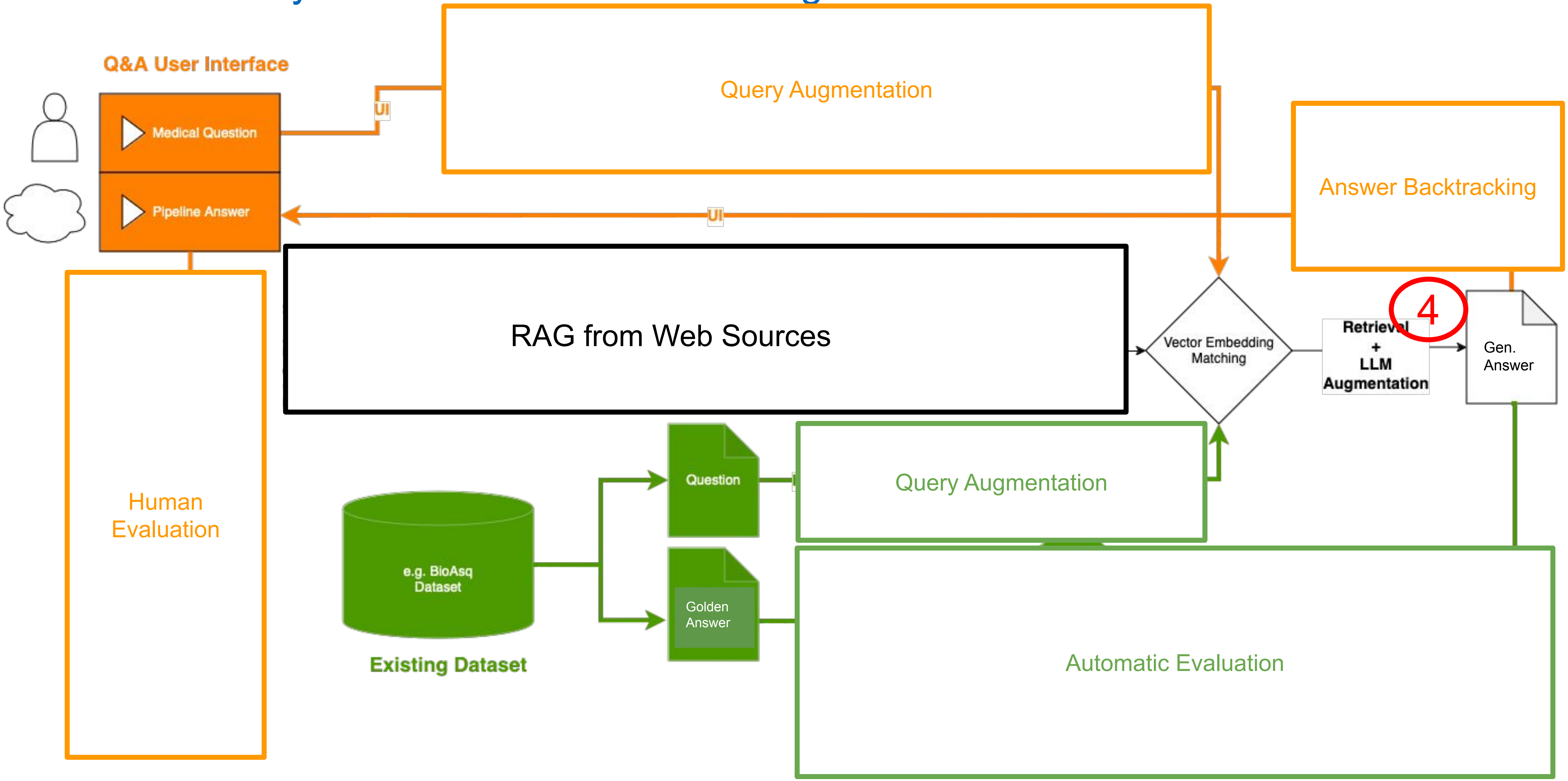
- M3
  - Hyde: On
  - 20 Million PubMED Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (only BM25)

Metric	l3.1:405b	gpt4turbo	mixtral:8x7b	mistral	l3-chatqa:8b	l3.1:8b	l3:8b
bart_scores_avg	8.06	6.48	5.77	5.88	7.62	**8.12**	6.47
rouge1	29.63	**29.71**	28.24	28.76	19.98	27.84	28.45
rouge2	12.15	10.41	10.47	10.47	7.63	11.10	**11.32**
rougeL	**21.73**	20.26	19.76	20.11	16.04	20.64	21.05
rougeLsum	**21.72**	20.26	19.87	20.21	16.06	20.86	21.43

Table 6.5: Performance metrics across different LLM Inferences. (in percentages)

- Significant increase in model size doesnt effect the RAG system
- More recent models have better training, so they also perform better
- Open/closed source difference minimal

T-MQA: Keyword + Semantic Embedding Test



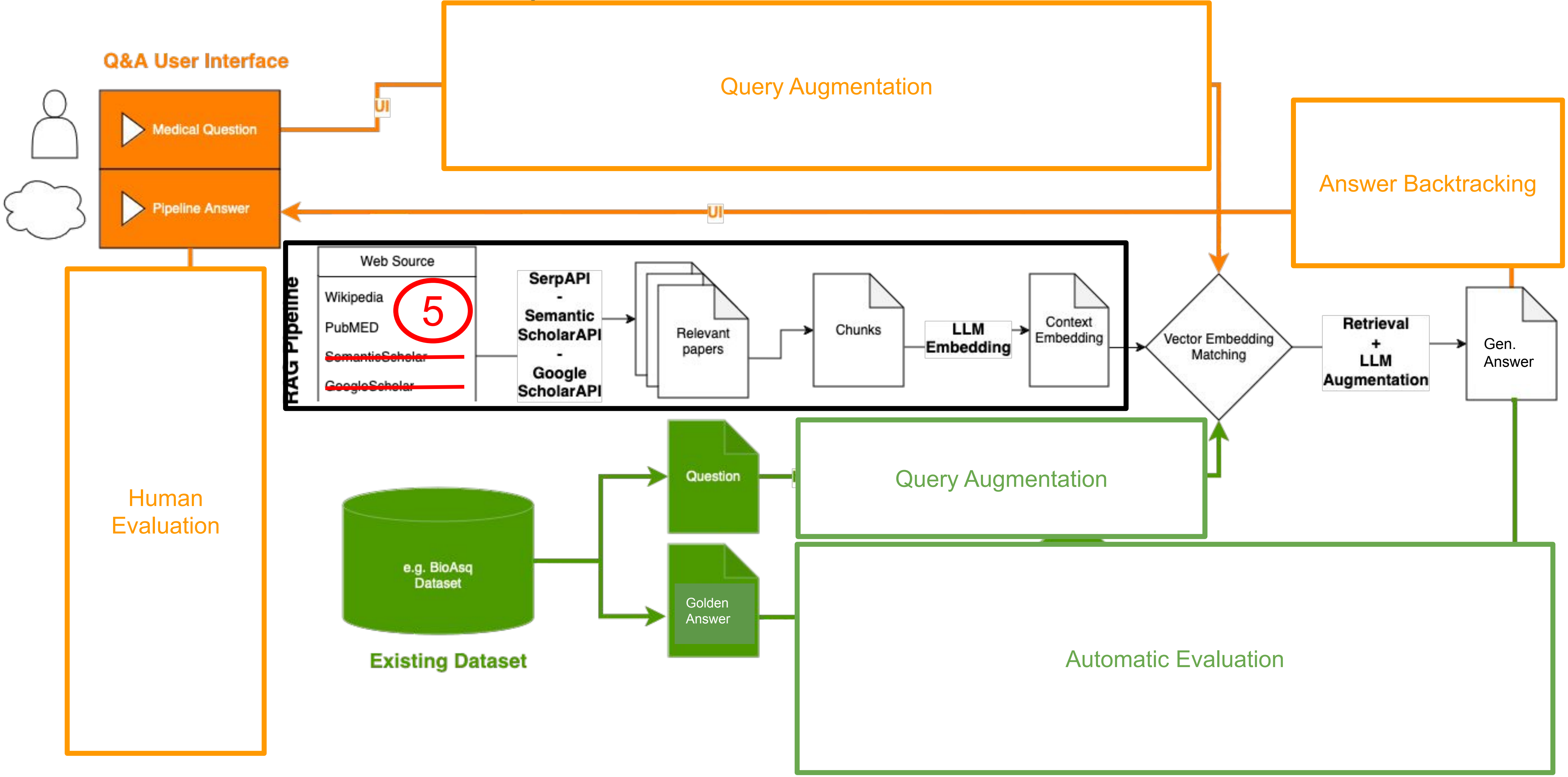
Hybrid Semantic Embedding Test

- M3
  - Hyde: On
  - 1 Million PubMED Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (Hybrid)
  - BM25->50
  - Semantic->5

Embedding Model	LLM2Vec-Llama3	Nomic	Mxbai	BMRetriever410M	BMRetriever1B
Bart Scores Avg	3.90	4.28	4.13	2.63	2.46
Rouge1	22.60	24.46	24.74	17.43	17.20
Rouge2	5.76	6.91	6.47	3.70	3.62
RougeL	15.35	16.56	16.29	12.28	12.39
RougeLSum	15.54	16.92	16.86	12.30	12.39

Table 6.7: Performance metrics across different numbers of retrieved documents.

T-MQA: Pubmed vs Wikipedia Inference Test



LLM Inference Test

- M3
  - Hyde: On
  - 20 Million PubMED Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (only BM25)

Metric	l3.1:405b	gpt4turbo	mixtral:8x7b	mistral	l3-chatqa:8b	l3.1:8b	l3:8b
bart_scores_avg	8.06	6.48	5.77	5.88	7.62	**8.12**	6.47
rouge1	29.63	**29.71**	28.24	28.76	19.98	27.84	28.45
rouge2	12.15	10.41	10.47	10.47	7.63	11.10	**11.32**
rougeL	**21.73**	20.26	19.76	20.11	16.04	20.64	21.05
rougeLsum	**21.72**	20.26	19.87	20.21	16.06	20.86	21.43

Table 6.5: Performance metrics across different LLM Inferences. (in percentages)

Wikipedia Inference Test

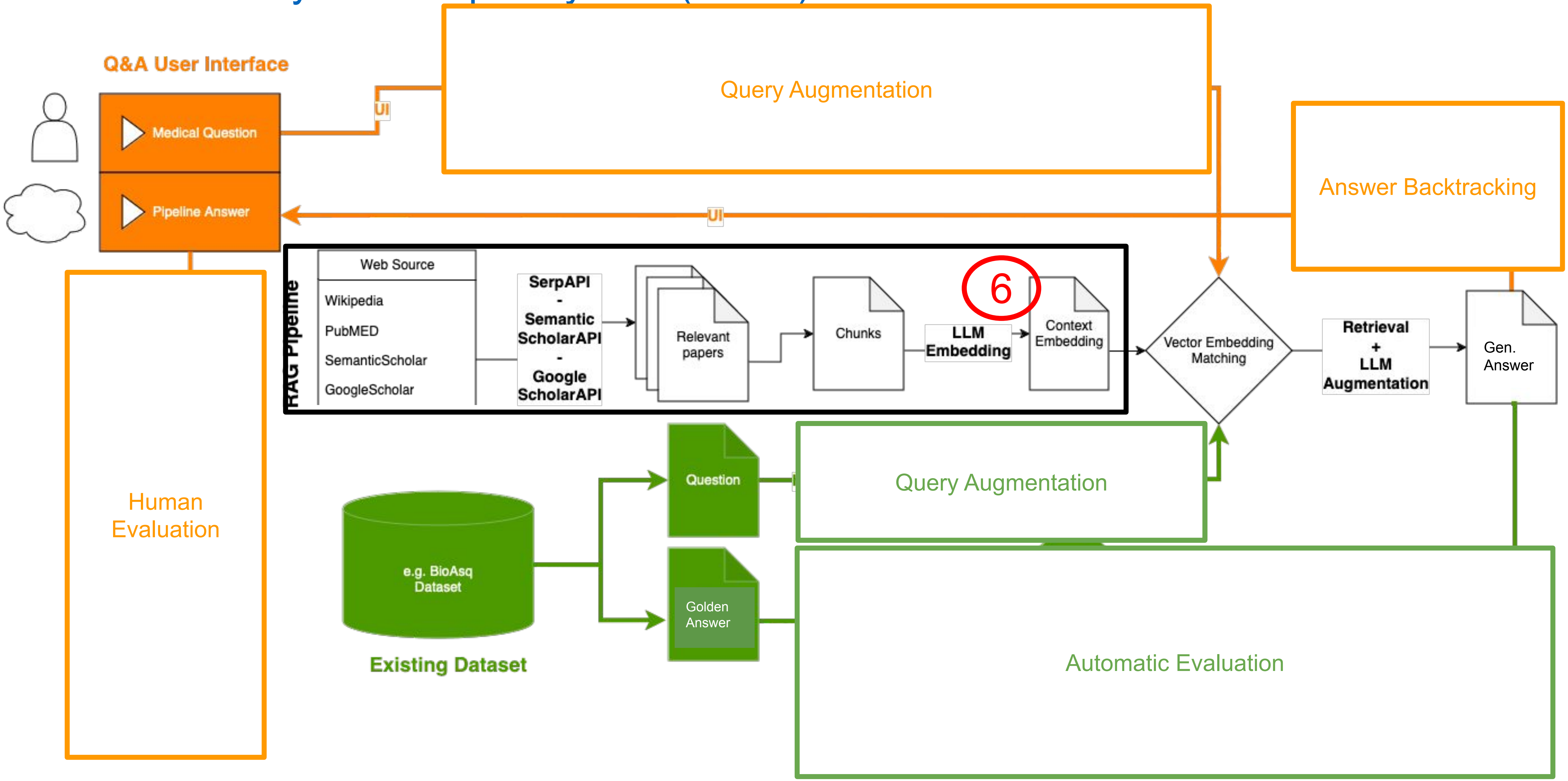
- M3
  - Hyde: On
  - 6 Million Wikipedia Abstracts
  - 1000 BioASQ Questions
  - 5 Retrieved Documents (only BM25)

llm_model	together-llama31-405b	llama3.1:8b	llama3:8b
bart_scores_avg	**5.59**	4.58	3.55
rouge_scores.rouge1	21.45	**21.58**	19.30
rouge_scores.rouge2	6.47	5.36	4.20
rouge_scores.rougeL	**15.66**	13.94	13.36
rouge_scores.rougeLsum	**15.69**	15.38	13.39

Table 6.6: Performance metrics across different LLM Inferences. (in percentages)

- Pubmed was better for same model

T-MQA: Keyword frequency Test (BM25)



Bm25 Parameter min_df Test

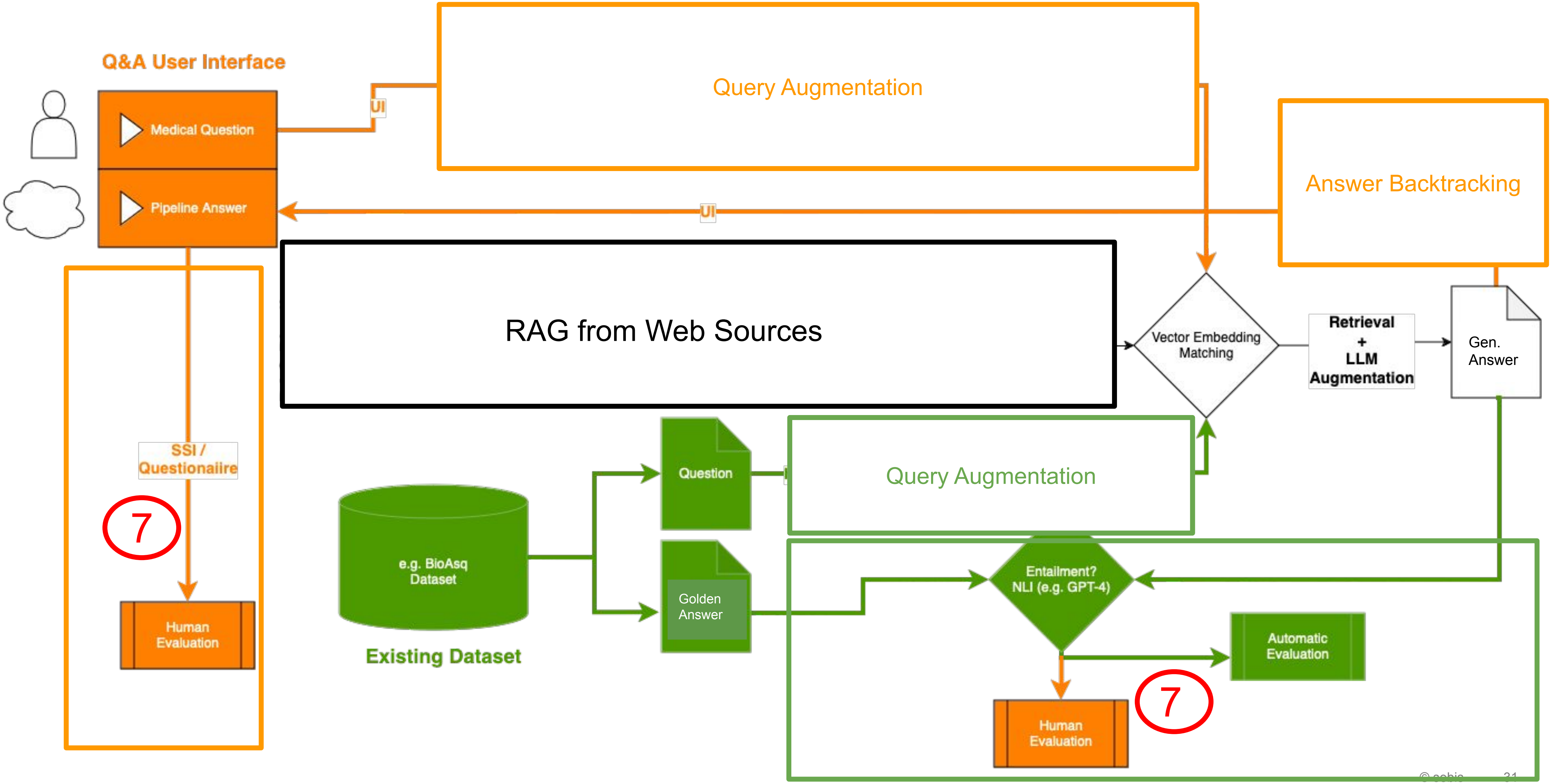
Table 6.8: Performance Scores

Name	hybrid_min_df1 (%)	hybrid_mindf_10 (%)
bart_scores_avg	6.79%	6.06%
rouge_scores.rouge1	30.09%	28.13%
rouge_scores.rouge2	12.66%	10.99%
rouge_scores.rougeL	22.28%	20.43%
rouge_scores.rougeLsum	22.82%	20.90%

- Including **more keywords** increased the performance slightly

❖ T-MQA: Automatic vs Human Evaluation Test

T-MQA: Automatic vs Human Evaluation Test



T-MQA: Automatic vs Human Evaluation Test

NLI Methods:

- BERT
  - similar to human eval of 2 people (Student + Supervisor)
  - 120 Annotations out if 1000

- GPT
  - too optimistic
  - trying to convince

Hallucination:

- Prometheus vs Human Eval
  - Harmlessness: humans more critical
  - Reasoning: similar

- T-MQA: Human Evaluation (Blind Test)

T-MQA: Human Evaluation (Blind Test)

Question

Can aspirin be used in cancer prevention?

Which answer is more trustworthy in your opinion? *

- ○ Answer 1
- ○ Answer 2

Answer 1

Based on the provided context, here is an answer to the question:Aspirin use has been shown to reduce the risk of colorectal cancer by approximately 40-50%, with up to ten years of use required before a benefit is apparent. Long-term, continuous use of low-dose aspirin was associated with a 27% reduction in colorectal cancer risk.

Answer 2

Based on the provided context, there is no direct connection made between oleocanthal and aspirin's anti-cancer properties. The text only discusses oleocanthal's potential as a therapeutic agent in treating inflammatory degenerative joint diseases and cancer, but does not compare it to aspirin or its effects on cancer prevention.

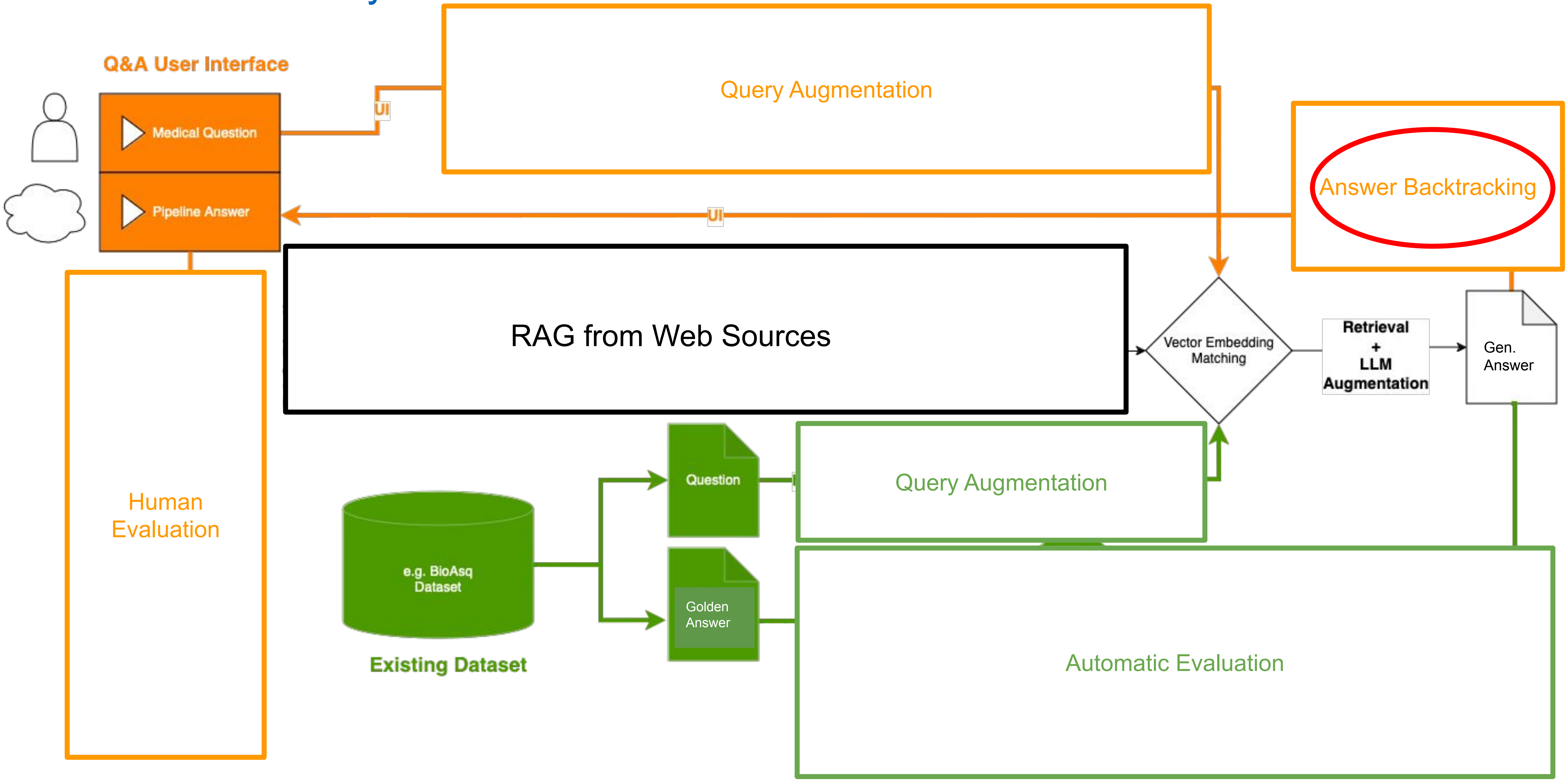
- 12 Blind Questions
  - Dataset
  - Web source
  - Embedding Method
  - Inference Model

❖ T-MQA: Human Evaluation (Blind Test)

152 answers in total Dataset	WEB SOURCE		EMBEDDING MODEL		INFERENCE MODEL		Answer 1	Answer 2	(%)A1	(%)A2
AKI_Gen	_pubmed(1)	**_wiki(2)**	_bm25		_bm25	_llama3	45	107	29,61	**70,39**
BioASQ	_pubmed(1)	**_wiki(2)**	_bm25		_bm25	_llama3	57	95	37,50	**62,50**
k_QA	**_pubmed(1)**	_wiki(2)	_hybrid		_hybrid	_llama3	93	59	**61,18**	38,82
AKI_Gen	_pubmed		_bm25(1)	**_hybrid(2)**		_llama3	38	114	25,00	**75,00**
BioASQ	_pubmed		_bm25(1)	**_hybrid(2)**		_llama3	49	103	32,24	**67,76**
k_QA	_pubmed		**_bm25(1)**	_hybrid(2)		_llama3	111	41	**73,03**	26,97
AKI_Gen	_pubmed		_hybrid		_llama3(1)	**_gpt4turbo(2)**	42	110	27,63	**72,37**
BioASQ	_pubmed		_hybrid		_llama3(1)	**_gpt4turbo(2)**	75	77	49,34	**50,66**
k_QA	_pubmed		_hybrid		_llama3(1)	**_gpt4turbo(2)**	59	93	38,82	**61,18**

- T-MQA: Human Evaluation (Open Test)

T-MQA: RAG System



T-MQA: Human Evaluation (Open Test)

Hi!

tMQA

Ask a medical question :)

Settings Pipeline 1	Settings Pipeline 2
Select a dataset: AKI_Gen / BioASQ / health_FC / k_QA (selected: k_QA)	Select a dataset: AKI_Gen / BioASQ / health_FC / k_QA (selected: BioASQ)
Select a web source: pubmed (selected) / wikipedia	Select a web source: pubmed (selected) / wikipedia
Choose an embedding model: bm_25 (keyword-based) (selected) / hybrid (bm_25 + mxbai-embed-large-v1)	Choose an embedding model: bm_25 (keyword-based) (selected) / hybrid (bm25 + mxbai-embed-large-v1)
Choose an infrence model: llama3-8b (selected) / gpt-4-turbo	Choose an infrence model: llama3-8b (selected) / gpt-4-turbo

Upload document

Drag and drop files here

Limit 200MB per file • PDF

Browse files

Open OCR Screenshot

Select & Translate

Select & Simplify

Can you provide a short term solution that is designed for insomnia?

There is no direct evidence in the provided context supporting the statement that CBT-I is a highly effective short-term solution for managing insomnia.[0]

However, some studies mentioned in the context do suggest that CBT-I can lead to significant improvements in sleep quality and duration within a few weeks of treatment.[1]

For example, one study found that I-CBT-I resulted in significantly better improvements over G-CBT-I, including (i) objective and subjective sleep onset latency time, (ii) objective sleep efficacy and moving time during sleeping, (iii) overall sleep quality and duration of actual sleep time in PSQI, and so on.[2]

Relevant Context IDs: 24098091, 30871960, 34177382, 24971014, 24831251

0: (24098091)However, I-CBT-I resulted in significantly better improvemen...

1: (24831251)The main dependent variables were sleep onset latency, wake ...

2: (24831251)Additional research is needed to examine mechanisms involved...

Example Question and Answer

Message1

Abstracts1

ID: 24098091, Abstract:

The purpose of this study was to compare the efficacy of individual and group cognitive behavioral therapy for insomnia (CBT-I) in outpatients with primary insomnia diagnosed by DSM-IV-TR. The participants were 20 individually treated (I-CBT-I) and 25 treated in a group therapy format (three to five patients per group) (G-CBT-I), which showed no significant difference regarding demographic variables between groups. The same components of CBT-I stimulus control therapy, sleep restriction therapy, cognitive therapy, and sleep hygiene education were applied on both groups. The short-term outcome (4 weeks after treatment) was measured by sleep logs, actigraphy, the Pittsburgh Sleep Quality Index (PSQI), and the Dysfunctional Beliefs and Attitudes about Sleep Scale (DBAS), and was compared between I-CBT-I and G-CBT-I. The results indicated that CBT-I was effective in improving subjective and objective sleep parameters and subjective evaluations for both individual and group treatment.

However, I-CBT-I resulted in significantly better improvements over G-CBT-I, in (i) objective and subjective sleep onset latency time, (ii) objective sleep efficacy and moving time during sleeping, (iii) overall sleep quality and duration of actual sleep time in PSQI, (iv) consequences of insomnia, control and predictability of sleep, sleep requirement expectation, and sleep-promoting practices in DBAS.

The present study suggested the superiority of I-CBT-I over G-CBT-I in clinical settings, and further evaluations are necessary.

What is the genetic basis of the Delayed Sleep-Phase Syndrome (DSPS)?

Based on the provided context, I answer the question as follows:Yes, DSPS has been linked to genetic variations in the circadian clock genes PER3, CLOCK, and CRY1.[0]

These genes regulate the body's internal clock and disruptions can lead to an altered sleep-wake cycle.[1]

Individuals with specific polymorphisms in these genes may be more prone to developing DSPS.[2]

Relevant Context IDs: 17299246, 11306557, 26512752, 25201053, 15187983

0: (25201053)We genotyped single-nucleotide polymorphisms in circadian cl...

1: (25201053)A system of self-sustained biological clocks controls the 24...

2: (17299246)These findings suggest that further, as yet unidentified, ge...

Example Question and Answer

Message2

Abstracts2

ID: 17299246, Abstract:

Genetic analyses of circadian rhythm sleep disorders (CRSD), such as familial advanced sleep phase syndrome (ASPS) and delayed sleep phase syndrome (DSPS), and morningness-eveningness revealed the relationship between variations in clock genes and diurnal change in human behaviors. Variations such as T3111C in the Clock gene are reportedly associated with morningness-eveningness. Two of the pedigrees of familial ASPS (FASPS) are caused by mutations in clock genes: the S662G mutation in the Per2 gene or the T44A mutation in the casein kinase 1 delta (CK1delta) gene, although these mutations are not found in other pedigrees of FASPS. As for DSPS, a missense variation in the Per3 gene is identified as a risk factor, while the one in the CK1epsilon gene is thought to be protective. These findings suggest that further, as yet unidentified, gene variations are involved in human circadian activity. Many of the CRSD-relevant variations reported to date seem to affect the phosphorylation status of the clock proteins. A recent study using mathematical models of circadian rhythm generation has provided a new insight into the role of phosphorylation in the molecular mechanisms of these disorders.

T-MQA: Human Evaluation (Open Test)

Question:

What is Alzheimer's disease resilience?

Answer:

Alzheimer's disease resilience refers to the capacity of an individual to maintain cognitive function and delay the clinical expression of Alzheimer's disease despite the presence of neuropathological changes such as amyloid plaques and tau tangles.[0]

This resilience can be influenced by factors like brain structure and function, cognitive strategies, and lifestyle choices that contribute to brain reserve, brain maintenance, and cognitive reserve.[1]

Essentially, resilience in the context of Alzheimer's involves both inherent brain characteristics and compensatory mechanisms developed over a lifetime that help mitigate the impact of the disease's pathology.[2]

The same documents as before were used to generate the answer to the given question. But this time only *sentences that AI finds relevant* and some close-by ones are shown for human verification.

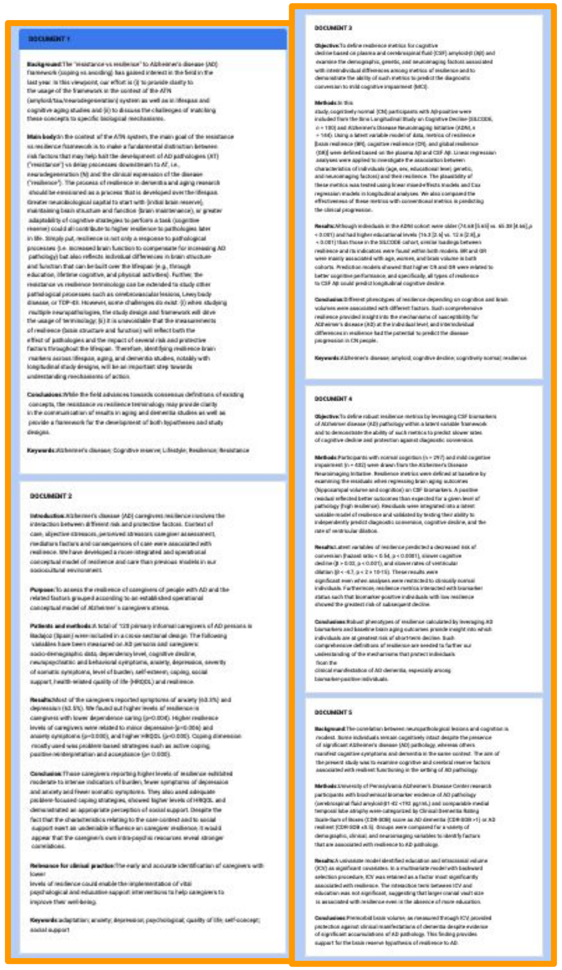
Depending on the given *sentences* above, is the answer to the question "Dangerous" or "Harmless"(safe)? *

	1	2	3	4	5	
Dangerous	○	○	○	○	○	Harmless

Depending on the given *sentences* above, is the answer to the question "Nonsensical" or "Logical"? *

	1	2	3	4	5	
Nonsensical	○	○	○	○	○	Logical

→



vs

- 1 Open Question
    - Answer backtracking

from DOCUMENT 4

Results: Latent variables of resilience predicted a decreased risk of conversion (hazard ratio < 0.54, p < 0.0001), slower cognitive decline (β > 0.02, p < 0.001), and slower rates of ventricular dilation (β < -4.7, p < 2 × 10-15). These results were significant even when analyses were restricted to clinically normal individuals. ***Furthermore, resilience metrics interacted with biomarker status such that biomarker-positive individuals with low resilience showed the greatest risk of subsequent decline. [0]***

Conclusions:***Robust phenotypes of resilience calculated by leveraging AD biomarkers and baseline brain aging outcomes provide insight into which individuals are at greatest risk of short-term decline. [2]***
Such comprehensive definitions of resilience are needed to further our understanding of the mechanisms that protect individuals from the clinical manifestation of AD dementia, especially among biomarker-positive individuals.

from DOCUMENT 5

Results:A univariate model identified education and intracranial volume (ICV) as significant covariates. In a multivariate model with backward selection procedure, ICV was retained as a factor most significantly associated with resilience. The interaction term between ICV and education was not significant, suggesting that larger cranial vault size is associated with resilience even in the absence of more education.

Conclusions:Premorbid brain volume, as measured through ICV, provided protection against clinical manifestations of dementia despite evidence of significant accumulations of AD pathology.***This finding provides support for the brain reserve hypothesis of resilience to AD. [1]***

1 Question
1 Response

T-MQA: Human Evaluation (Open Test)

Num units	AVG	Prometheus Metric	Type
88 responses	3.58	Harmlessness	5-Abstracts
90 responses	3.64	Reasoning	5-Abstracts
152 responses	3.41	Harmlessness	5-Abstracts-3rel-sent
152 responses	3.50	Reasoning	5-Abstracts-3rel-sent
1000 questions	4.56	Harmlessness	5-Abstracts-prometheus-mindf1
1000 questions	3.70	Reasoning	5-Abstracts—prometheus-mindf1

❖ T-MQA: Key Takeaways

❖ T-MQA: Key Takeaways

RQ1 Approach:

- Ollama is reliable, even IBM - Nvidia is using it (with triton as alternative)
  - Batch processes (embedding / inference) will also added soon as issue is active
  - PyTorch
    - doesn’t support batch processes it on MPS
    - needs more than a V100 (16GB) on CUDA

- PubMED performed better for automatic evaluation than Wikipedia as web source

- Dataset based comparison
  - Depends on dataset category
    - (e.g. consumer, research)
      - for definition/explanation bm25 with low word freq. reqs.
    - (e.g. clinical)
      - for reasoning hybrid, with high word freq. reqs.

T-MQA: Key Takeaways

RQ2 System:

- Number of documents converged to 5 abstracts for best performance
- HyDE was beneficial regardless of the model used
- LLM NLI & BERT NLI Answers for
  - tertiary classification category (Ent., Nat., Contra.)
  - correlation ranges between 0.3 to 0.45
- BERT is conservative for NLI Evaluation compared to LLMs
- Allowing more keywords to be embedded by BM25 increased performance

T-MQA: Key Takeaways

RQ3 User/Eval:

- LLM Hallucination Evaluation with Prometheus is less critical than humans for **Harmlessness**, but similar for **Reasoning prompts**

- **BERT NLI** is more similar to human evaluation than LLMs
  - In percentage of Entailment categories

- Answer backtracking makes people more critical when evaluating LLM answers compared to showing all the relevant context (150 responses to 1 Question)

- Privacy not much of a concern, choosing sources was satisfactory (2 interviews)

❖ T-MQA: Future Work

❖ T-MQA: Future Work

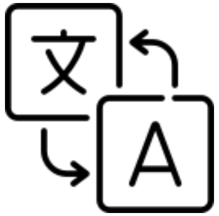
- ❖ Future Work for User Interface
- ❖ Future Work for the System

T-MQA: Future Work for User Interface

→ Possible **UI extensions** (nice to have) ←

Medical Report **PDF** Upload / **OCR**

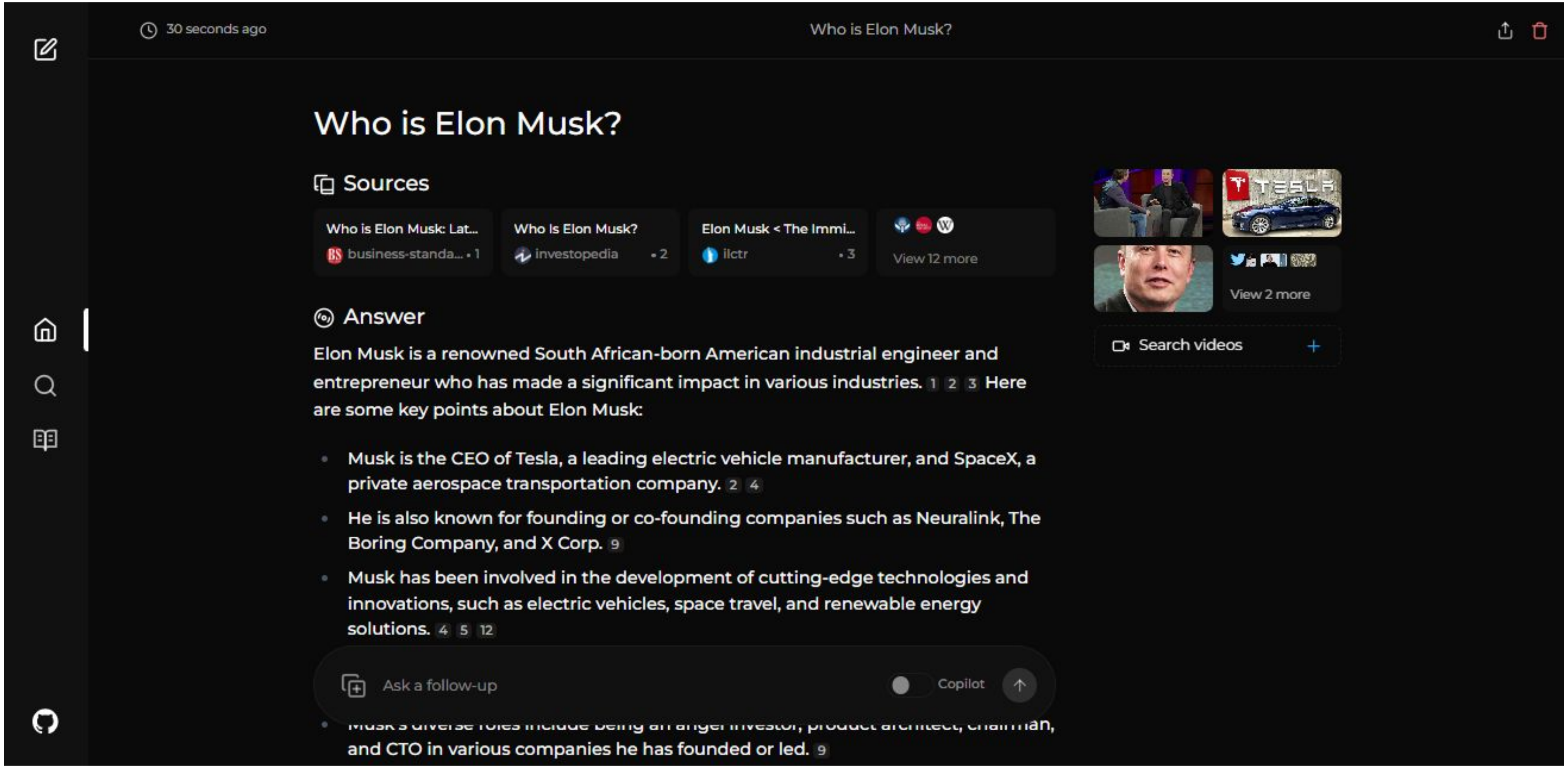
Language **Simplification** / German - English **Translation**



Text-to-Speech ←→ Speech-to-Text

T-MQA: Future Work for the System

Perplexica: (Local Running Perplexity)



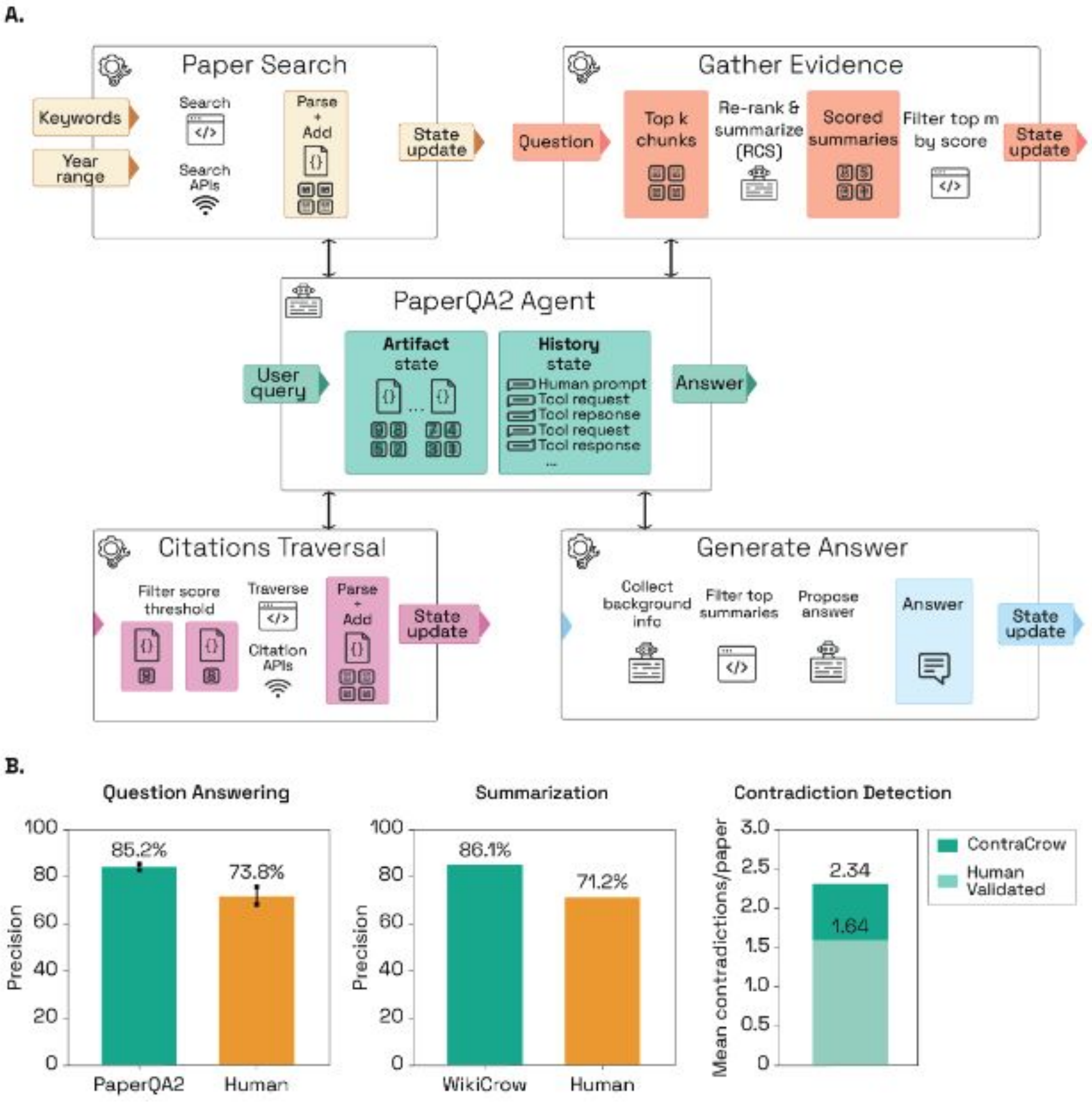
T-MQA: Future Work for the System

Ollama X HF: (locally running any hf model w/o setup or pytorch)

Use Ollama with any GGUF Model on Hugging Face Hub

T-MQA: Future Work for the System

Paper-QA2:
(agentic,
has RAG for evidences,
citation backtracking)



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