DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in TUM School of Computation, Information and Technology (CIT): Robotics, Cognition, Intelligence

Developing Systems for Trustworthy Medical Question Answering

B.Sc. Ragip Volkan Tatlikazan

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in TUM School of Computation, Information and Technology (CIT): Robotics, Cognition, Intelligence

Developing Systems for Trustworthy Medical Question Answering

Entwicklung von Systemen zur vertrauenswürdigen Beantwortung medizinischer Fragen

Author:	B.Sc. Ragip Volkan Tatlikazan
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Juraj Vladika
Submission Date:	10.10.2024

I confirm that this master's thesis in tum school of computation, information and technology (cit): robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Munich, 10.10.2024 B.Sc. Ragip Volkan Tatlikazan

1 AI Assistant Usage Disclosure

1.1 Introduction

Performing work or conducting research at the Chair of Software Engineering for Business Information Systems (sebis) at TUM often entails dynamic and multi-faceted tasks. At sebis, we promote the responsible use of AI Assistants in the effective and efficient completion of such work. However, in the spirit of ethical and transparent research, we require all student researchers working with sebis to disclose their usage of such assistants. For examples of correct and incorrect AI Assistant usage, please refer to the original, unabridged version of this form, located at this link.

1.2 Use of AI Assistants for Research Purposes

I have used AI Assistant(s) for the purposes of my research as part of this thesis.

[X] Yes [ ] No

Explanation: After writing the Introduction, Conclusion, Abstract and Acknowledgements myself, I used ChatGPT to improve those sections in the following areas: language and style enhancement, clarity and conciseness. Then I used it to translate the abstract to the German language. Additionally, I used ChatGPT for transforming some citation strings into bibtex entries. Finally, I used Grammarly's browser extension to identify mistakes, typos or unclear sentences in the whole work. The license for Grammarly was provided by TUM. I confirm in signing below, that I have reported all usage of AI Assistants for my research, and that the report is truthful and complete.

Munich, 10.10.2024 B.Sc. Ragip Volkan Tatlikazan

Acknowledgments

I'm grateful to everyone who have made this research possible and enriched the experience. First, I would like to thank M.Sc. Juraj Vladika and Prof. Dr. Florian Matthes for their guidance and support throughout this project. Their expertise and insights have been fundamental to my research.

I am also grateful to M.Sc. Nicolas Berberich for his guidance and perspectives which have led me to this work.

Special thanks are due to Murilo Bellatini, Alexander Blatzheim, and Jonas Lossin for their previous work.

I am also very grateful to my family and friends for their unwavering support and encouragement. Their belief in my work has been a constant source of motivation.

I appreciate the collective wisdom and contributions of the broader research community, whose work has laid the groundwork for and inspired my own.

Finally, I would like to acknowledge the financial support of German taxpayers. Their contributions made it possible for me to study in Germany.

Abstract

Since several years AI has been on the rise, and specifically in the field of NLP, innovation is very fast-paced. These advances bring tremendous benefits to humanity but they also come with a lot of risks.

The speed of AI in combination with its confident tone in its oratory while hallucinating factual information is clearly dangerous. One of the areas that it can be the most dangerous to humanity is medicine.

With the current trajectory of humanity's adaptation to chat interfaces, makes it important to make developments in the area of trustworthy and reliable medical information distribution.

This study focuses on developing systems for trustworthy medical question answering. We explore various aspects and methodologies in the context of medical question answering systems. We investigated the optimal number of documents retrieved, comparing performance when varying the count, and found that five documents ensure a safer information scope than three, despite the latter showing slightly better performance. Our testing on different hardware, including V100 GPUs and cloud services like Google Cloud Platform and Google Colab as well as a M3 Max, indicates that satisfactory results can still be achieved without top-tier equipment.

The retrieval efficiency of the BM25 algorithm was highlighted as it outperformed the combination of semantic vector-based search and BM25 in a hybrid manner by a factor of ten in setup, with modifications in keyword frequency thresholds enhancing performance due to the unique nature of medical terminology. Human evaluations and annotations pointed out the critical role of the LLMs wording in the absence of context and its relative influence next to different BERT-based NLI models, which were the closest to human intuition.

Additionally, we analyzed the performance impact of different Large Language Models (LLMs) and query expansion method HyDE, where newer LLM versions did not significantly outperform older ones despite substantial size differences. In contrast, modifications in document count and LLM type used for inference showed notable improvements. Furthermore, hybrid retrieval methods combining BM25 with semantic re-ranking models demonstrated a quantitative decrease in performance, emphasizing the necessity of optimized embedding strategies for efficient implementation.

These findings suggest careful consideration of hardware, retrieval techniques, and LLM capabilities in developing effective medical question answering systems, while not refuting their value in their current form. It can already increase the understanding and efficiency of interacting with medical information in a private and trustworthy manner.

Kurzfassung

Seit einigen Jahren ist Künstliche Intelligenz insbesondere im Bereich der Sprachverarbeitung (NLP) auf dem Vormarsch. Diese Fortschritte bringen enorme Vorteile für die Menschheit, aber die Geschwindigkeit der KI in Kombination mit ihrem selbstsicheren Ton, während sie faktisch falsche Informationen liefert, ist eindeutig gefährlich. Ein Bereich, in dem sie für die Menschheit am gefährlichsten sein könnte, ist die Medizin.

Angesichts der aktuellen Anpassung der Menschheit an Chat-Schnittstellen ist es wichtig, Entwicklungen im Bereich der vertrauenswürdigen und zuverlässigen Verteilung medizinischer Informationen voranzutreiben.

In dieser Studie konzentrieren wir uns auf die Entwicklung von Systemen für vertrauenswürdiges medizinisches Frage-Antworten. Wir erforschen verschiedene Aspekte und Methoden im Kontext von medizinischen Frage-Antwort-Systemen. Wir haben die optimale Anzahl von abgerufenen Dokumenten untersucht und festgestellt, dass fünf Dokumente einen sichereren Informationsbereich als drei gewährleisten, obwohl letztere leicht bessere Leistung zeigten. Unsere Tests auf unterschiedlicher Hardware, einschließlich V100-GPUs und Cloud-Diensten wie Google Cloud Platform und Google Colab sowie einem M3 Max, zeigen, dass zufriedenstellende Ergebnisse auch ohne Spitzengeräte erzielt werden können.

Die Effizienz des BM25-Algorithmus wurde hervorgehoben, da er in einer hybriden Kombination aus semantischer vektorbasierter Suche und BM25 um das Zehnfache in der Einrichtung übertraf, wobei Modifikationen in der Schlüsselwortfrequenzschwelle die Leistung aufgrund der einzigartigen Natur der medizinischen Terminologie steigerten. Menschliche Bewertungen und Annotationen zeigten die kritische Rolle der Formulierungen der LLMs bei Fehlen von Kontext und ihre relative Bedeutung neben verschiedenen BERT-basierten NLI-Modellen, die der menschlichen Intuition am nächsten kamen.

Zusätzlich analysierten wir die Leistungsauswirkungen verschiedener großer Sprachmodelle (LLMs) und der Abfrageerweiterungsmethode HyDE, wobei neuere LLM-Versionen trotz erheblicher Größenunterschiede ältere Versionen nicht signifikant übertrafen. Im Gegensatz dazu zeigten Modifikationen in der Dokumentanzahl und der für die Inferenz verwendeten LLM-Typen bemerkenswerte Verbesserungen. Darüber hinaus zeigten hybride Abfragemethoden, die BM25 mit semantischen Neurangierungsmodellen kombinieren, einen quantitativen Rückgang der Leistung, was die Notwendigkeit optimierter Einbettungsstrategien für eine effiziente Implementierung betont.

Diese Erkenntnisse legen nahe, dass sorgfältige Überlegungen zu Hardware, Abruftechniken und LLM-Fähigkeiten bei der Entwicklung effektiver medizinischer Frage-Antwort-Systeme erforderlich sind, während ihre Verwendung in ihrer aktuellen Form nicht abgelehnt wird. Sie kann bereits das Verständnis und die Effizienz der Interaktion mit medizinischen Informationen auf private und vertrauenswürdige Weise verbessern.

Contents

1 Introduction

1.1 Motivation

The major data leaks happening over the last 10 years from big tech companies, along with developments in AI—especially the impact of models like ChatGPT—have made the need for more secure and private medical information systems very clear. These developments, combined with the exponentially increasing number of medical research articles and the help of AI in all areas of research, have motivated my master's thesis on trustworthy medical question answering for patients, medical students, researchers, and doctors.

Since the beginning of 2024, I have been working on a hopefully to be open-source, locally running application that allows users to interact with their chosen source of medical information in privacy. This app can function without an internet connection once the necessary data and AI models are installed locally. Users can select from sources such as PubMed articles, Wikipedia, or any PDF of their choosing, similar to the functionality offered by Perplexica on GitHub. An another alternative can be the recent work PaperQA.

The project operates under several assumptions:

- Patients prefer receiving answers from reliable sources like scientific articles rather than from blog posts, newspapers, or acquaintances.
- Due to time constraints, patients often do not want to read lengthy scientific articles.
- The complex medical language in scientific literature and the typical short duration of medical consultations—often limited to 15 minutes—meaning that patients do not fully understand their diagnoses.
- Doctors must stay current with new developments in their field despite working long hours, including night shifts. Researchers often conduct redundant studies because it is virtually impossible to scan millions of articles each year before selecting a research topic.

To address these challenges and improve trust and understanding in medical processes, I have made my application modular for easy testing and clarity. It is designed to provide users the freedom to select their preferred sources of medical information, and in doing so enhancing the usability and adaptability of the system to individual needs.

*Data leaks: *Statista Cyber Crime*

**from PubMED (y-axis in Millions):

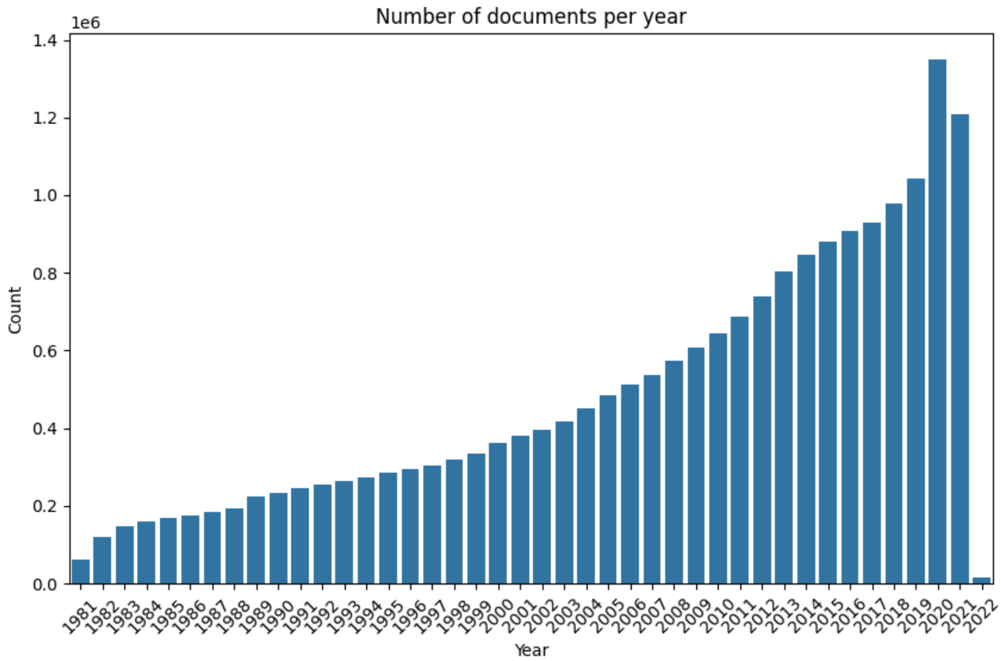

Figure 1.1: PubMED: Number of publications in provided Dataset by Year

1.2 Problem Statement

1.3 Research Questions (RQs)

The goal of this research is to develop systems that provide trustworthy answers that are needed in medical situations. For this, we go after the following 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 generate answers to medical questions using retrieved medical evidence (or knowledge) using LLMs and methods like RAG (Retrieval-augmented generation)?

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

1.4 Outline

The outline of the thesis is structured as follows.

Chapter 2 Fundamentals, goes through to fundamental methods in NLP used in the thesis.

Chapter 3 Related Work, points to the research articles this work is in general related to exploring the possibilities for RQ1 as widely as possible, only a subset of these articles are used in this work.

Chapter 4 Methodology, explains how we have combined the methods we chose to use from "Related Work" chapter to answer RQ2.

Chapter 5 Experiments, goes through which experiments we chose to do on the developed pipeline for RQ2.

Chapter 6 Results, is the aggregation of the evaluation results from the quantitative experiments and human evaluation, ergo answering RQ3.

Chapter 7 Discussion, speculates on the the results and draws possible explanations from the evaluation.

Chapter 8 Conclusion, is a short summarization of all the chapters and the gives conclusion of the thesis.

2 Fundamentals

2.1 NLP

2.1.1 Natural Language Processing

Nature Language Processing is a part of AI that specializes in textual information. It can be divided into subresearch areas being text comprehension, text summarization, text generation, and text classification. An NLP software can do this by dividing the textual information down to paragraphs, sentences, and then into words which are called tokens in this area.

NLP has been used in translating languages, formatting text in recent years and in generating answers to questions on existing data. The foundational steps of NLP start with tokenization. Afterwards, it was be extended to removal of unnecessary words, identification of parts of speech in sentences, utilizing methods like bag of words for categorization and vectorization. In recent years, more increasingly in sentence vectorization, starting with the sequential deep learning methods, recurrent neural networks and long short-term memory networks(LSTMs) and recently after the breakthrough with transformers utilizing the attention concept, processing in a non-sequential way as well.

For more detailed information please refer to the publication "A Survey of the Usages of Deep Learning for Natural Language Processing" [1]

2.1.2 Attention

The attention mechanism allows the models to dynamically focus on the contextually relevant parts of the input text. Initially used to improve translation, now it makes it possible to capture the importance of every word in a sentence. It can identify the contextually most relevant words in a sentence and numerically give less significance to words which are more frequent and don't carry significant information. This method has been so successful that it is widely adopted for almost all NLP tasks.

This attention mechanism is mainly used in a transformer layer in neural networks, which has several self-attention heads that allow the model to capture different aspects of textual formation. For those interested, you can find the full calculation of the Transformer layer in the Appendix. For the inner workings of the transformer layer, there are numerous sources online.

2.1.3 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. BERT goes through words in sentences, in relation to all the other words simultaneously. It does not work sequentially like RNNs. This allows BERT to make better connections between words than sequential methods, as sentences can be constructed in a one-directional way. As it is parameter-wise a relatively small type of model, it can be fine-tuned easily with current technology on specific tasks very easily such as text classification or Seq2Seq models for translation. It is also widely used to learn domain-specific information too. For more detailed information on BERT please refer to the original paper. [2]

2.2 LLM: Large Language Models

Another currently very popular and healthy model is the Large Language Models (LLMs) like GPT (Generative Pre-trained Transformer). These are very advanced AI models that can comprehend entire human languages and deep linguistic patterns by essentially predicting the next word sequentially after being trained off Basically the whole internet. They can also textually manipulate intricasies in languages such as idioms, grammar, and even sarcasm. They are highly capable for a range of applications, including conversational agents, content creation and so on.

2.2.1 RTE: Recognizing Textual Entailment

RTE is a linguistic task where a statement in one piece of text is evaluated computationally to understand if it logically follows from another. These two pieces of text are called hypothesis and premise in order. This evaluation results in entailment and non-entailment.

It is a cornerstone NLP task that requires deep knowledge of semantics, information extraction, and summarization. The biggest problems to tackle in RTE are faced when processing expressions which have very similar meanings or when paraphrasing.

2.2.2 NLI

Natural language inference is the updated version of RTE which also includes a neutral option as a result of the evaluation, in addition to the entailment and contradiction. It results in a neutral evaluation when the message of a premise neither confirms or refuses the message in the hypothesis. And a contradiction happens when the premise suggests that the hypothesis is false.

2.2.3 RAG: Retrieval Augmented Generation

RAG is a very important advancement in NLP to solve the biggest problem of hallucination that the large language models face one trying to answer questions truthfully. It does this by using information retrieval systems which are then fed into the large language model when a question is being asked. This way a large language model can depend on the given context when inferring an answer with its linguistic knowledge. while ignoring contradicting facts that are simultaneously stored in the LLM and then, moving to the training phase, resulting in reliable answers with higher readability that are based on real-world data.

2.2.4 Quantization

Quantization is also an important concept in the context of LLMs. It refers to the method that reduces the numerical parameter precision so that it becomes smaller to store computationally. Typically, the elements are trained with floating-point numbers. However, after the training, a big part of the decimals can be disregarded, resulting in a significantly decreased on the size of the model. which also increases the speed of 3 operations done with the LLM. Quantization is crucial for deploying LLMs on devices such as mobile phones or embedded systems.

3 Related Work

3.1 Coding Language

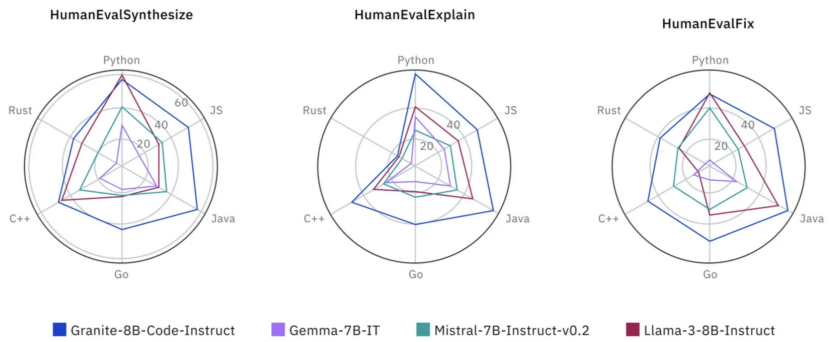

Figure 3.1: IBM Granite for major languages Task-Coding language Performance evaluation

Given the recent advancements in large language models, we have chosen to go forward with the coding language python for this project, for the simplicity of further development. As you can see in the figure, IBM Granite for major languages [3], the potential of synthesizing and explaining code is maximized in Python. Assuming the trend of using LLMs to code continue, we predict that the usage of a more general language would contribute more in total.

3.2 Surveys

The following two surveys have been the starting point of this research. To get started with biomedical question answering or fact-checking. Please give these two papers a short look.

3.2.1 Biomedical Question Answering: A Survey of Approaches and Challenges

The first survey has also been a good categorization and inspiration, for the datasets, which in turn accelerated the choice of format to interact with biomedical data. [4]

3.2.2 Scientific Fact-Checking: A Survey of Resources and Approaches

The second survey provided the initial dataset to base the search for fitting datasets for the task as you can see on Figure 3.2 and a base for a pipeline that we decided on, to pursue medical question answering. Although it comes from a claim verification research. You can see the example in the following Figure 3.3. [5]

Dataset	# Claims	Claim Origin	Evidence Source	Domain
SciFact (Wadden et al., 2020)	1,409	Researchers	Research papers	Biomedical
PubHealth (Kotonya and Toni, 2020b)	11,832	Fact-checkers	Fact-checking sites	Public health
Climate-FEVER (Diggelmann et al., 2020)	1,535	News articles	Wikipedia articles	Climate change
HealthVer (Sarrouti et al., 2021)	1,855	Search queries	Research papers	Health
COVID-Fact (Saakyan et al., 2021)	4,086	Reddit posts	Research, news	COVID-19
CoVERT (Mohr et al., 2022)	300	Twitter posts	Research, news	Biomedical

Table 1: Datasets for the task of scientific fact-checking and claim verification

Figure 3.2: Scientific Fact Checking Datasets

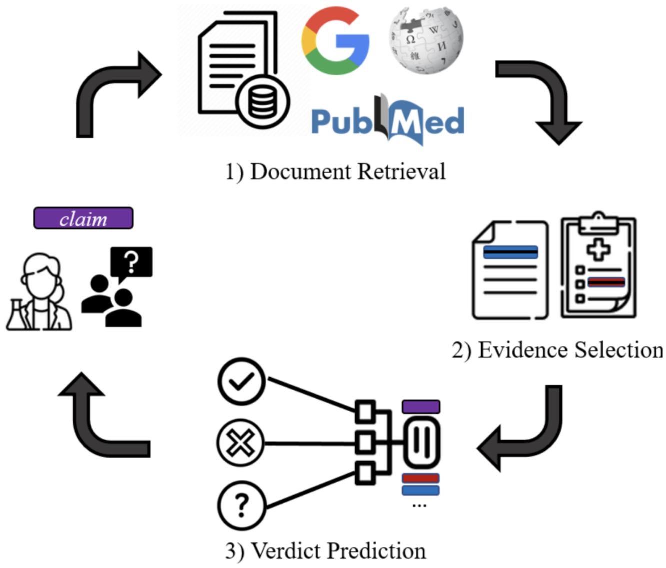

Figure 3.3: Base for Proposed Pipeline

3.3 Datasets

In this section. you can see the aggregation of several datasets for medical question answering or claim verification. We use the same categorization from the biomedical question and answering survey. The categories are research, consumer, clinical, and examination questions or claims. We also include the year of the publication regarding the given datasets and how many citings they had on 2024 April for credibility.

The check-marked columns are the features of the given datasets, addition to the questions. Most of them also have an ideal answer and context to reach that answer. We also hypothesize that a question-and-answer pair is equivalent to a claim-and-label pair. However, this hypothesis hasn't been used in our tests. additional augmentation and a check for semantic equivalence of these two pairs would be needed and can be an additional module to the pipeline.

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	✓			✓			

Table 3.1: Medical QA Datasets

As you can see from the table, the BioASQ is the most detailed dataset to the best of our knowledge. Here you can see in Figure 3.4 an example of its features, the types of the values of those features, and a given example. We make references to this structure when possible to provide an easier understanding of the other datasets structures.

Below is an analysis of the dataset papers that can be related for the medical question and answering tasks in the same order. Additionally, there is a short explanation of the features for most of the datasets to save time for a future work in this topic.

For an easier read, you can think of context as a paragraph given to answer the question and evidence as the part of the paragraph that is the most relevant to the answer. The term document is used to refer to the whole data structure or to the context most of the time.

- BioASQ 11b Dataset [7]

  BioASQ 11b dataset is from 2023 and has several previous versions in previous years. It has IDs for the questions and answers, paper snippets, the links for the full papers, and additional question types which are "yes/no", "factoid, "list" or "summary". Factoid means short sentence as an answer.

- SciFact Dataset [8]

  SciFact dataset is from 2020 with 336 citations. It has a claim, evidence for the claim, and binary classification labels in general, that are relevant. Its features also include an

Filed	Type	Content
id	String	A unique identifier of the question. E.g. "52bf1b0a03868f1b06000009"
body	String	The question body in English. E.g. "What is the mode of inheritance of Wilson's disease (WD)?"
type	String	The question type in English. One of "yesno", "factoid", "list" or "summary"
documents	Array of Strings	List of relevant article URLs. E.g. ["https://www.ncbi.nlm.nih.gov/pubmed/838566",…]
snippets	Array of JSON Objects	List of relevant snippets. E.g. [{"offsetInBeginSection":122, "offsetInEndSection":272, "text":"The disease…", "beginSection":"abstract", "document":"http:…", "endSection":"abstract"},…]
concepts	Array of Strings	List of relevant concept URLs. E.g. ["https://www.disease-ontology.org/api/metadata/DOID:893",…]
triples	Array of JSON Objects	List of relevant triples. E.g. [{"p":"http:…/name", "s":"http:…/diseases/1198", "o":"Wilson_disease"},…]
ideal_answer	Array of Strings	List of ideal answers to the question in English. E.g. ["WD is an autosomal recessive disorder.",…]
exact_answer not available in summary questions	Depends on the type of the question	For *yesno*: A String ("yes" or "no") For *factoid*: An array of Strings, synonyms of the answer. E.g. ["CaM kinase II", "CAMK2"] For *list*: An array of arrays of Strings with synonyms of each element of the answer. E.g. [["Triadin","TrD"], ["Calsequestrin", "CASQ",…],…]

Figure 3.4: BioASQ Features [6]

ID, broken-down version of the context into sentences and the citations. It also has a big corpus which includes document IDs, titles, and how the corpus is structured.

- HealthVER Dataset [9]

  HealthVER dataset is from 2021 and has been cited 28 times. You: It has a claim, evidence, and binary classification, similar to the SciFact dataset. The dataset additionally has a tertiary label, a topic IP (which is unmentioned in the paper nor the code), and additionally the question. The claim column represents an answer.

- PubHealth Dataset [10]

  PubHealth dataset is from 2020 and has been cited 168 times. It has a claim, a label, and an explanation, similar to the Bio-Ask 6B and 7B. This dataset has four different labels, which are true, false, mixture, and unproven.

  This dataset is similar to the BioASQ 11b in the sense that it is very detailed. In addition to the above, it has the publishing dates of the evidence text, the fact checkers and the authors, the source of the evidence, a subject for all of these claims to categorize them easier, sources for the claims, and the name of the websites usable for categorizing as well.

- MEDIQA-AnS [11]

  MEDIQA-AnS dataset is from 2020 with 65 citings. Besides the usual question answer and several documents, the *question type* explains the medical application as well, such as interaction, information, usage/time, dose, and side effects. It also includes the rating for the answers from 1 to 4. (e.g. "3-Incomplete", "4-Excellent") Metadata LINK

- MedQuAD [12]

  MedQuAD Dataset is from 2019 and has been cited 48 times. Dataset columns are a document, question, and answer. And additionally, there is a *question type* that gives the category of the medical area such as treatment, info, cause, and so on.

- LiveQA [13]

  LiveQA Dataset is from 2017 with 64 citings. It has a question and answer, which are manually retrieved by librarians. It also has a *question type* which gives the category of the medical area similar to the MedQuAD dataset. (treatment, info, cause)

- K-QA [14]

  K-QA Dataset is from 2024, cited only one time, but has the important distinction of separating the answer into sentences and labeling the must-have sentences and nice-to-have sentences depending on the question. Of course, it also has the question and answer but the source of the context is not given as text but only as a link.

- CliCR Dataset [15]

  CliCR Dataset is from 2018 and has been cited 95 times. The dataset is structured in the way that there is a document that has the context and the title of the evidence with its source and then there is the question and answer which also has the origin of the answer. Data structure GitHub LINK Different from other basic sets, this also includes a semantic type of the answer. An example for that could be a patient problem and diagnosis and so on. The dataset is available per email request.

- emrQA [16]

  emrQA Dataset is from 2018 and has been cited 185 times. It has the usual features, question, answer, general context, and the evidence part from it and additionally an answer entity type. The *answer entity type* is given as empty, single, or complex which is basically means there is no answer, only one piece of evidence, or several pieces of evidence.

- HEAD-QA [17]

  HEAD-QA Dataset is from 2019 and has been cited 45 times. It has multiple choice questions and answers, which also include question categories, such as medicine, nursing, psychology, chemistry, pharmacology, and biology.

- HealthFC [18]

  HealthFC dataset is from 2024 has been cited 2 times. It has the claim, context, evidence, and a tertiary label constructed of, supported, not enough information, and refuted. The dataset is available both in English and German. Also the date of the articles.

- AKI_Gen (Alpha KI Generated) [19]

Alpha KI project is a digital health assistant project from the SEBIS chair at TUM. It is backed by the Bavarian Ministry of Economic Affairs and Regional Development and Energy. It has been active since 2021 and it ends at 2024. One of the student theses that contributed to this project was on generating question-and-answer pairs as a dataset to utilize for generative AI. This dataset has 3500 questions answered pairs and 200 of them were reviewed by medical practitioners. We utilize this dataset as well to show the performance differences on human-created datasets and generated datasets.

3.4 General (non-medical) QA Datasets

This section has general, non-medical question-answering datasets to compare the differences between these and medical datasets.

- HotpotQA [20]

  HotpotQA Dataset is from 2018 and has been cited 1,721 times. It has the usual question, answer, context, and evidence features. Additionally, the *question type* is given as comparison or bridge and the *question level* being easy, medium, or hard.

- TriviaQA [21]

  TriviaQA Dataset is from 2017 and has been cited 1,768 times. It has the usual columns question, answer, and context. It additionally includes the *source of the question* and the *results of the web search to give the answer*.

- SQuAD [22]

  SQuAD Dataset is from 2016 and has been selected 7,759 times. It has the usual question answer and the context included. It also has a detailed linguistic analysis.

- Unanswerable Questions for SQuAD [23]

  This dataset is from 2018 and has been cited 2682 times. The structure is quite similar to the squad dataset, but the idea of retraining with unanswerable questions of the given dataset is to emphasize here. As it might improve the performance in total and show the shortcomings of the first model in the medical domain as well.

3.5 Web Sources

- Wikipedia Wikipedia Plaintext (2023-07-01)

  The Wikipedia web source given has the following features. The id, the title, the first sentence of the article as context, an additional column called text, which is the concatenation of title and context and the category of the given Wikipedia article.

- Wikipedia Full Text Plain Text Wikipedia 2020-11

The Wikipedia full-text web source given has only the title, full text, and the id. An updated Wikipedia dump can be processed by the given link to include the category as well similar to the above data structure if needed. (Wikipedia dump handling code)

- PubMED [24]

  Dataset Link: The landscape of biomedical research dataset

  The PubMed web source given has the article abstract, the title, the ID, and the publication year. It also has the journal the article belongs to. The link also shows PubMed embeddings with float 16 precision.

  The paper also provides the unsupervised clustering method t-SNE with embedding X and Y coordinates, labels, and color for a better understanding of the biomedical research area distribution. You can see the general mapping here, but the paper also has additional sub-mappings, which are worthwhile to look at. As you can see from the image it is easy to denote that biomedical research has many domains that can benefit from a specialized language model.

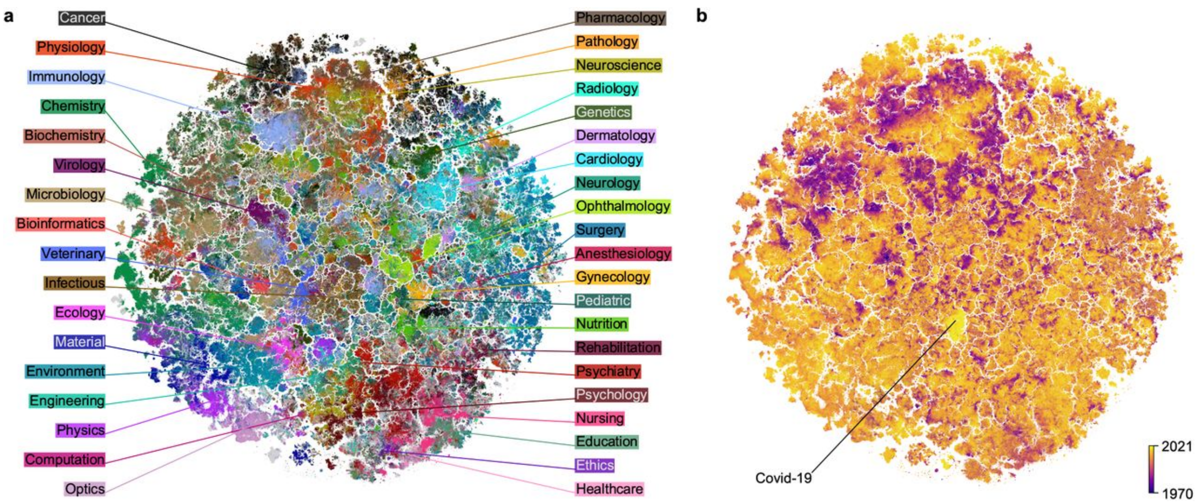

Figure 3.5: t-SNE PubMED Landscape

  The provided PubMedBERT embeddings with float16 precision are not adaptable to our use case. As PubMedBERT embeddings don't utilize the Siamese-BERT (sBERT [25]), which are very high in performance for sentence matching compared to the regular BERT embeddings.

- k-QA List of Sources [14]

  Another important figure to inspect is the list of web sources used from the K-QA paper. It provides a good starting point for online sources of medical information. They list 15 different most used web sources. by six medical physicians for 201 questions. To make it easier for medical personnel, we try to keep similar sources in this work. "uptodate.com" is a subscription-based service. "my.clevelandclinic.org" and "mayoclinic.org" don't include sources, they do provide a nice health library for definitions but instead of this, a general anatomy book in addition to a pathalogy book would be a good enough replacement if needed. As they can cover the same scope and adding more official medical books instead of unreferenced online sources would be more reliable. "ncbi.mlm.nih.gov", "medlineplus.gov", "pubmed.ncbi.mlm.nih.gov", "emedicine.medscape.com" have references and sources.

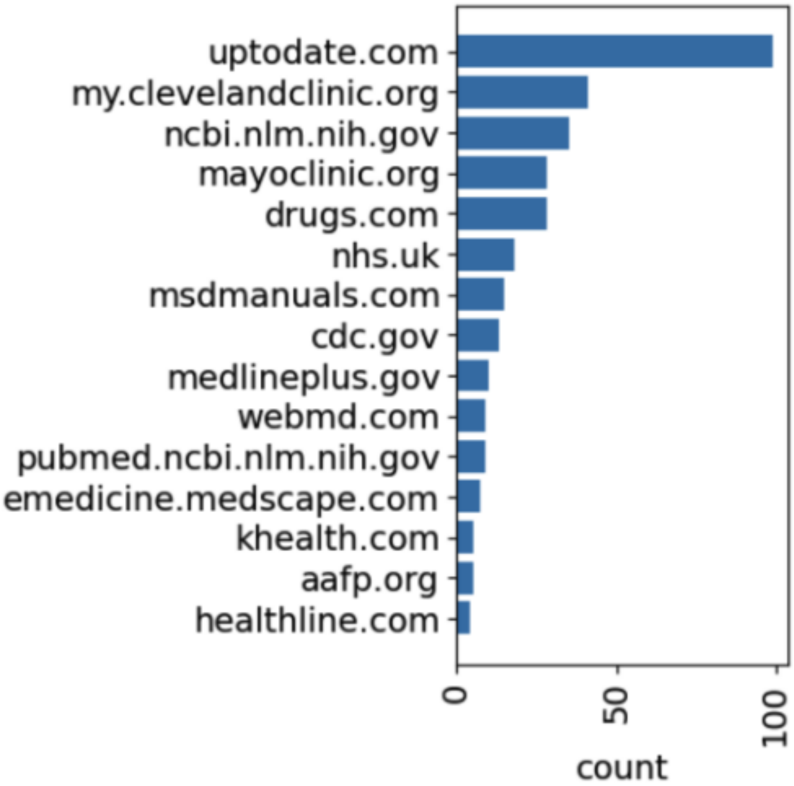

Figure 3.6: k-QA List of Sources

- Other Sources Other possible sources can be Google Scholar and Semantic Scholar. Semantic Scholar has an API that is an academic Graph dependent on the SPECTER2 embedding. They also have an API for their Datasets curated from the papers, and Peer Review API for "Detection of conflict of interest, based on co-author relationships", "Computation of a matching score between a reviewer and a submission's topic, based on the reviewer's publication history". Their "ask this paper" AI feature is a good example of a retrieval system based on a question by a user. It is only available on a subset of the papers they provide access, but the selection is unclear. You can find more information on SPECTER2 in SciRepEval Paper. [26]

  Google Scholar can be used with Serp API. Serp API can also query a normal Google search for medical information. This way a lot of additional metadata can be used to filter through medical research papers. Such as the year of the publication, number of citations, related articles that side this paper.and of course the authors.

3.6 Benchmark Papers

In this section, you can find two papers for benchmarking LLMs used in medical domain.

Model	*Comp* ↑	*Hall* ↓	% respond
MedAlpaca 7B	31.4	56.7	100
Mistral 7B	47.6	28.4	100
PALM-2	50.8	31.3	100
BARD	62.5	28.4	95.0
Bing Chat	57.3	25.9	99.5
GPT-3.5	56.2	27.9	100
GPT-3.5+ICL	59.5	23.4	99.5
GPT-3.5+RAG	50.5	17.9	89.0
GPT-3.5+ICL+RAG	62.9	**15.4**	96.0
GPT-4	57.5	23.9	100
GPT-4+ICL	**67.7**	25.4	100
GPT-4+RAG	52.2	22.9	91.5
GPT-4+ICL+RAG	65.2	24.4	100

Table 4: Comparing models, where ICL represents the addition of three in-context examples, and RAG is the medical retrieval augmented setup, as detailed in Section 5.1. The performance of the highest scoring model is **bolded** for each metric. *% respond* indicates the percentage of generations that do not abstain from answering the questions.

Figure 3.7: k-QA Model Comparison

Figure 3.7 shows several models including in-context learning and retrieval augmented generation methods. What's important to recognize here is that hallucinations go down when both ICL(in context learning) and RAG is used. [14]

		MCQA	AQA		
		Acc	RL	BS	MTR
Base	LLaMA 2 (7B)	42.9	14.9	55.3	21.1
	LLaMA 2 (13B)	47.1	15.0	56.4	22.5
	MPT (7B)	27.6	13.3	52.6	21.1
	Falcon (7B)	34.7	14.0	54.1	20.0
Instruction tuned	LLaMA 2-chat (7B)	45.9	15.0	58.0	23.3
	LLaMA 2-chat (13B)	50.3	15.3	58.0	23.6
	MPT-Instruct (7B)	31.6	15.8	59.7	15.6
	Falcon-Instruct (7B)	31.8	17.2	62.4	17.4
	Flan-T5 (3B)	51.8	10.8	55.0	7.4
	Flan-T5 (11B)	56.5	11.5	56.3	8.2
Adopted	ChatDoctor (7B)	42.8	17.4	62.3	18.7
	MedAlpaca (7B)	48.8	15.5	58.9	15.6
	PMC-LLama (13B)	53.7	19.7	60.7	19.0

Table 3: Zero-shot performance of base (top), instruction-tuned models (middle) and domain-adopted (bottom) models. Metrics are **Acc**uracy for MCQA; **R**ouge-**L**, **BERTS**core, and **METEOR** for AQA.

Figure 3.8: M-QALM Benchmarks

Abstract question answering (AQA) means synthesizing a answer from given sources instead of choosing from a pre-existing text. Figure 3.7 shows that in the AQA category domain-adopted models doesn't include a significant increase in performance regarding Rouge-L, BERTscore and METEOR scoring methods. [27]

3.7 Method Papers

Table 6. Systems and approaches for task 10b. Systems for which no information was available at the time of writing are omitted.

Systems	Phase	Approach
bio-answerfinder	A, B	Bio-AnswerFinder, ElasticSearch, Bio-ELECTRA, ELECTRA, BioBERT, SQuAD, wRWMD, BM25, LSTM, T5
bioinfo	A, B	BM25, ElasticSearch, distant learning, DeepRank, universal weighting passage mechanism (UPWM), PARADE-CNN, PubMedBERT
LaRSA	A, B	ElasticSearch, BM25, SQuAD, Marco Passage Ranking, BioBERT, BoolQA, BART
ELECTROBERT	A, B	ELECTRA, ALBERT, BioELECTRA, BERT
RYGH	A	BM25, BioBERT, PubMedBERT, T5, BERTMeSH, SciBERT
gsl	A	BM25, BERT, dual-encoder
BioNIR	A	sBERT, distance metrics
KU-systems	B	BioBERT, data augmentation
MQ	B	tf-idf, sBERT, DistilBERT
Ir_sys	B	BERT, SQuAD1.0, SpanBERT, XLNet, PubMedBERT, BioELECTRA, BioALBERT, BART
UDEL-LAB	B	BioM-ALBERT, BioM-ELECTRA, SQuAD
MQU	B	BART, summarization
NCU-IISR/AS-GIS	B	BioBERT, BERTScore, SQuAD, logistic-regression

Figure 3.9: BioASQ Systems

[28] BioASQ being the most comprehensive dataset also has a reoccurring challenge, Where biomedical question and answering methods and systems are tested. In Figure 3.9 you can see several systems and which approaches they have used so far. There is heavy usage of pre-trained BERT models. ranking models, some traditional machine learning methods, and keyword search models.

[29] In this paper you can find one of the first context-based question-answering datasets that have been introduced to the field of NLP. It also includes the idea of the attention method compared to LSTM's two years before the heavily cited paper, "Attention is all you need".

[30] In this paper, you can see the now popularized method re-ranking of the retrieved information from a knowledge base and the hybrid usage of and regeneration. A visual aid is provided in Figure 3.10.

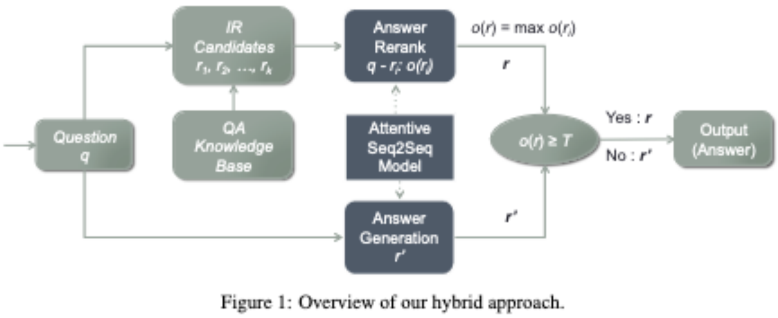

Figure 3.10: AliChat hybrid approach

[31] In this paper, you can see the previously mentioned hybrid method for information retrieval from Q&A knowledge bases applied to citations. The hybrid steps are more clearly defined for scientific purposes. These two are a selection of the candidate citations and then re-ranking them before providing it. A short diagram is provided in Figure 3.11.

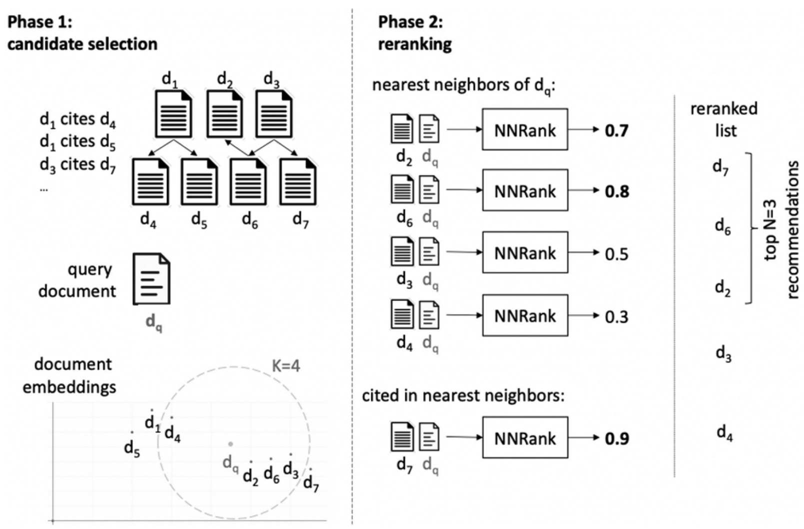

Figure 3.11: Hybrid citation recommendation

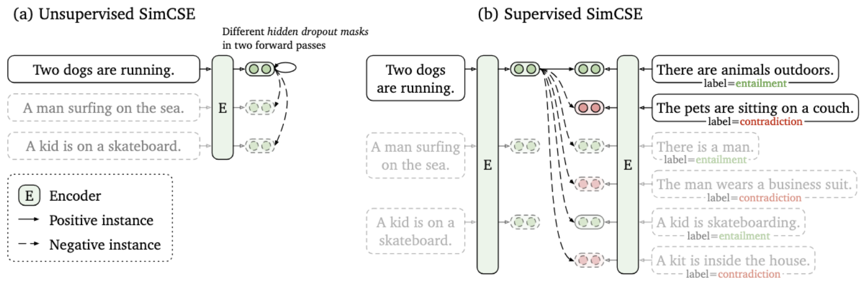

Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch instances as negatives.

Figure 3.12: Contrastive Learning

[32] This paper puts forward the sentence-based contrastive learning for embeddings, which means increasing the clarity between matched semantically sentences, by grouping the closer ones together and pushing the different ones further apart. This can provide a higher resolution of medical terms when it is added to the matchings between sentences of an LLM-generated answer and sentences of a given context. A detailed visualization can be found on the bigger 3.12.

[33] In this recent paper from 2024, you can see a good system architecture for claim verification. It provides a good system draft to transition into a question-answer-based architecture.

[34] Improving Health Question Answering with Reliable and Time-Aware Evidence Retrieval In this paper you can see experiments for number of retrieved documents, number of retrieved sentences, and different takes on the year of the retrieved documents were published. can provide valuable insight into our work as well.

[35] In this very recent paper to give attention to which is very similar and more comprehensive to this work is the BM Retriever. It has a comprehensive comparison of models that are around 1 billion parameter size and detailed aggregation of relevant Q&A datasets in biomedical domain.

3.8 LLM

[36] In this paper we have an open source pre-trained LLM for medical domain. We predict that the advance of open source LLMs will continue to improve with better distillation methods that are applied on bigger models. For this reason we believe that pretraining LLMs is not a logical way of developing sustainable systems. It is also important to mention that pre-training for domain specific tasks is useful, but is also makes the evaluation of the answer complicated, as tracing the reasoning of the given answer is not possible.

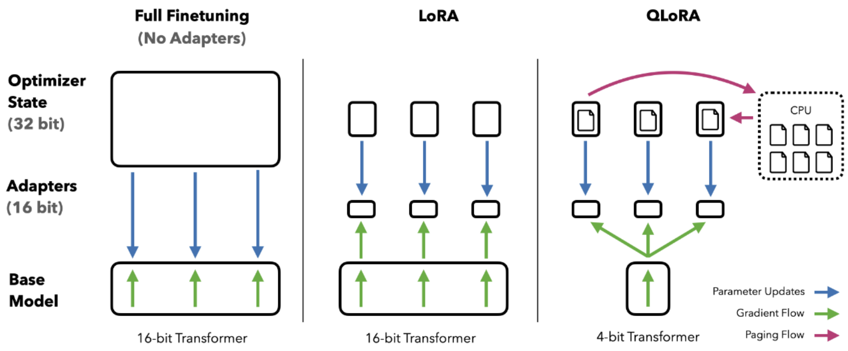

Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

Figure 3.13: Finetuning, LoRA, Q-LorA

[37]

If you are in need of a more closed system, which doesn't have many steps, and pre-training is the method you want to pursue, this paper, mentioning LORA and Q-LORA to make the fine-tuning efficient is a relevant to make your system adaptable to changes. You can see a visualization of the differences between the mentioned concepts in Figure 3.12

The key difference they have added is "injecting trainable rank decomposition matrices into each layer of the Transformer architecture" to reduce the amount of parameters to train drastically. Here is also the predecessor paper for more information: [38]

[39] In this recent paper, it has been suggested that LLMs can also be used at text encoders, besides the usual BERT size models. For the same reason, mentioned at the start of this section that the improvements in LLMs will continue. This might be an important method when implementing new systems. As it can half to time for searching an embedding model and inference model by basically using the same one and applying the trans formation of the LLM to an encoder.

3.9 Prompting

[40] In this paper you can find prompts to start with for several categories such as math, coding and natural language reasoning that incorporate chain of thought method.

[41] In this paper you can find prompts for evaluating a Q&A pair for reasoning, doing a claim verification check based only on given answer and predicting the next question, all of these steps can help with creating a highly relevant chat history and suggesting further questions.

[42] In this paper you can see a combination of first order logic and LLMs for explainable claim verification. Mainly the method proposed is to convert sentences into predicates and then using the LLMs power to do reasoning on this format.

[27] In this paper you can see a comparison of instruction tuned models and domain

adapted models for abstract question answering. The metrics they use for the evaluation are accuracy, Rogue-L, BERTScore and Meteor. These metrics are also interesting to look at for our case. They provide numerous prompts for one shot, few shot, single context, multi context, multiple choice and abstract question answering,

3.10 Document - Query Expansion

[43] In this paper you can find the query augmentation method which suggests changing the format of a question to a claim, which can provide a higher performance when doing semantic sentence matching, caused by similarity in formatting. In figure 3.14 you can see an example of how this transformation is done and additionally a quote from the paper itself for more detailed explanation.

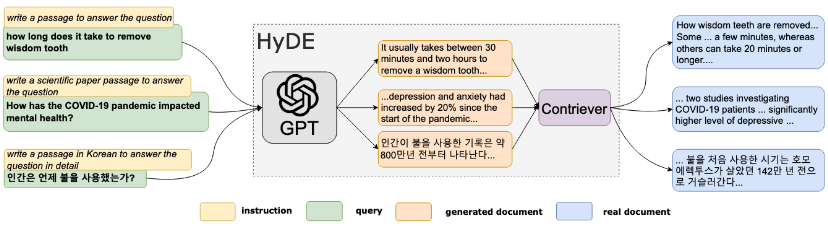

Figure 3.14: HyDE

"Hypothetical Document Embeddings (HyDE): Given a query, HyDE first zero-shot instructs an instruction-following language model (e.g. InstructGPT) to gen- erate a hypothetical document. The docu- ment captures relevance patterns but is unreal and may contain false details. Then, an un- supervised contrastively learned encoder (e.g. Contriever) encodes the document into an embedding vector."

[44] This paper suggests using the opposite direction as the above paper for bringing a query and a document together. They suggest predicting a query from a given document and adding this to the document itself to make the search more relevant for possible undefined queries. They use BM-25 as their baseline and expand on it.

[45] In this paper they focus on query and document expansions, specifically for domain, relevance and format shifts. Domain Shift refers to changing the usage of a model for example general to specific. Relevance Shift refers to changing the model usage from Searching for a topical relevance to find control arguments or refuting claims. Finally, format shift refers to changing the length of a given query or document. On table 7 of the paper you can see that the above mentioned method HyDE as a significant improvement for queries which is relevant for question answering.

Reranker vs Scoring agent

Retrieval metric	Context_precision		Context_recall	
(Number of document retrieved)	Reranker	Scoring agent	Reranker	Scoring agent
Question(12) + Hypothetical answer(12)	**0.717**	0.454	**0.328**	0.261
Multiquery questions(24)	**0.564**	0.36	0.269	**0.313**
HyDE with reranker/ScoringLLM (24)	**0.673**	0.43	0.283	**0.342**
Only question(24)	**0.556**	0.389	**0.27**	0.263
Only hypothetical answer(24)	**0.713**	0.41	**0.295**	0.279

Table 2: Comparison results of Reranker vs ScoringLLM

Figure 3.15: Hyde vs only hypothetical answer

[46] In this paper for using rag on a medical domain, they have incorporating a different approach in comparison with the HyDE method of using the hypothetical answer by a domain fine-tuned LLM for retrieval. They suggest that the hypothetical document generation given a query in the above mentioned HyDE paper from a general purpose LLM is very incomplete and it should be done with a domain-based fine-tuned LLM for this task. They show a non-negligible increase in the performance in their paper for the re-rankers, to be seen in figure 3.15

3.11 Automatic Evaluation

3.11.1 Metrics

[47] Rogue score is one of the most used automatic evaluation metrics in NLP. To shortly clarify the four different root scores from the original paper, you can read the following list:

- "rouge1": unigram (1-gram) based scoring
- "rouge2": bigram (2-gram) based scoring
- "rougeL": Longest common subsequence based scoring.
- "rougeLSum": splits text using "\n"

,with n-gram meaning "N-grams are sequences of characters or words extracted from a text. N-grams can be divided into twocategories: 1) character based and 2) word based." [48]

[49] BARTScore is a recent evaluation metric that is important for text generation tasks. To quote from the paper itself, "The basic idea is that models trained to convert generated text to/from a reference or source text will achieve higher scores for better quality generated text." In simpler terms, computing probabilities of generating one text given another.

[50] This paper puts forward the FACTSCORE. It uses atomic facts in a given document to evaluate it: Firstly, by separating the document into sentences and then those sentences into smaller sentences which contain one verifiable information and then evaluating them one by

one to give a total score on the document itself. This results in a more nuanced evaluation of a document.

3.11.2 Hallucinations

Another important topic for information retrieval is hallucinations. [51] In this paper they suggest forward the DeBERTa based model to evaluate if an LLM has generated centent that wasn't included in retrieved documents. Luna uses token level hallucination detection, that is then aggregated. over sentences for a total score.

[52] This paper suggests a broader approach for hallucination detection. Given the rack system, they aggregate the question context and the answer together and ask to the LLM Lynx for reasoning faults.

[53] In this paper they put forward and add a lamp with evaluation capabilities on par with GPT-4 named Prometheus. They also add the Quads structure from the system, which they share on GitHub as well. They utilize a scoring system from 1 to 5 and have a general prompt available with subcategories helpfulness, harmlessness, honesty, factual validity and reasoning, that they names rubrics.

Pearson vs Spearman vs Kendall's Tau Here you can find a quick guide for the correlation metrics.

3.12 Human Evaluation

Besides the automatic evaluation, we also have to consider human evaluation of LLM of generated text.

3.12.1 Question Categories

Type / Dataset	Question	Context	Answer
Scientific			
BioASQ	Is the protein Papilin secreted?	[...] secreted extracellular matrix proteins, mig-6/papilin [...]	Yes
Biomed-Cloze	Helicases are motor proteins that unwind double stranded _?_ into [...]	Defects in helicase function have been associated with [...]	nucleic acid
Clinical			
emrQA	Has the patient ever had an abnormal BMI?	08/31/96 [...] BMI: 33.4 Obese, high risk. Pulse: 60. resp. rate: 18	BMI: 33.4 Obese, high risk
CliCR	If steroids are used , great caution should be exercised on their gradual tapering to avoid _?_	[...] Thereafter, tapering of corticosteroids was initiated with no clinical relapse. [...]	relapse
Consumer			
MedQuAD	Who is at risk for Langerhans Cell Histiocytosis?	NA	Anything that increases your risk of [...]
MEDIQA-AnS	What is the consensus of medical doctors as to whether asthma can be cured? And do you have [...]	Asthma Overview Asthma is a chronic lung disease that causes episodes of wheezing [...]	Asthma is a chronic disease. This means that it can be treated but not cured. [...]
Examination			
HEAD-QA	The antibiotic treatment of choice for [...] is	1. Gentamicin; 2. Erythromycin; 3. Ciprofloxacin; 4. Cefotaxime	4. Cefotaxime

Table 2. Typical question-answer examples of different content types.

Figure 3.16: Typical question-answer examples of different content types[4]

A short look at the question categories is important here. As you can see from the Figure 3.15, we have scientific, clinical, consumer and examination questions.

[54] Each of these question categories correspond to a different way of the human ratings that they refer to in the G-Eval paper based on SummEval. These are coherence, consistency, fluency and relevance. For example, where a scientific answer has a higher importance in consistency, a consumer answer can have higher importance on fluency.

[55] In this paper, they put forward a language model called LIMA, which doesn't depend on reinforcement learning, but a very attentively constructed prompt and response pairs. To compare their language model with human responses, they propose the following human interface for annotations.

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.

Figure 11: Human annotation interface.

Figure 3.17: LIMA RLHF

4 Methodology

4.1 Scraping Data

Scraping data became irrelevant after the literature research as we want to keep our focus on running the entire system locally for privacy. But if needed, you can start by taking a look at SERP API or Semantic Scholar API. For a thesis focused more on this approach please search for "Analyzing and Improving Post-hoc Approaches for the Detection and Correction of Hallucinations in Long-form Text Generation"

4.2 Pre-processing

4.2.1 Datasets

The datasets we use are BioASQ 6b/7b/11b, HealthFC, the generated data set for Alpha KI project at SEBIS and KQA. The choices were made based on publication year, number of features for future work, availability of the dataset in Q&A format, and Q&A categories clinical, consumer, and research. We excluded education category for this project as it was based more on definition-based/multiple-choice questions and factual information that didn't have an interpretation aspect that LLM would face.

One quick note is that almost all features these datasets have are quite understandable, as seen in their respective papers from related work. Besides the "is-impossible" feature the BioASQ-11b dataset has.

It basically means that the question cannot be answered given the context in the dataset. Which can be seen as a third category added to the binary classification. (yes/no)

However, claim-based data sets shouldn't be neglected for future research. Using the claim and the label, we can achieve the same core information that a question-answer pair has. This approach is simultaneously quite similar to the idea that hypothetical document entailment paper shortly HyDE explores. So our methodology is also applicable to the claim-based data sets with an additional pre-processing step to turn these into question-answer pairs.

Given the power of LLMs to generate text, but also the increasing risk of running to hallucinations while generating longer amount of text like paragraphs, we decided to focus on "factoids" rather than the other categories that were presented in the most detailed dataset BioASQ which were "yes/no", "multiple choice" and "summary" question answer payers.

Additionally, you can see in the BioASQ systems overview figure in the related work, BERT architecture-based studies have been done extensively, which perform quite well for "yes/no" and "multiple choice" tasks.

4.2.2 Web Sources

We use two web sources, Wikipedia and PubMED. We chose these two for the availability they provided on Kaggle and Zenodo. Additionally, to be able to see the difference between a medical research paper web source and a general public information web source.

PubMed abstracts provided by Zenodo were fairly straightforward to process only needing a concatenation of the two different *csv* files for metadata/text, and sorting by year afterward.

The PubMedBERT embeddings that are provided on Zenodo are not usable for our case. As it is not a sBERT model, which is the Siamese network-based method that increases the performance of similarity search significantly. [25]

For Wikipedia, we concatenated the *JSON* files provided and then disregarded the Wikipedia pages that have more than 50,000 characters (26,630 out of 6,144,363) because of **pandas.DataFrame** cell variable type conversion limits.

4.3 Embedding

4.3.1 BERT

There have been many studies so far for the medical domain with BERT models that you can see on the BioASQ systems in related work You can look at the examples of domain-specific models which are the following: medicalBERT, distilBERT, pubmedBERT, BioBERT, and Clinical longformer. BERT models are quite small in size and comparison to LLMs which makes them versatile to train for specific tasks. But for our use case of a general system, we don't put an emphasis on methods for pre-training. So our design choices are based more on different types of embedding models, usability and sustainability.

4.3.2 Embedding models

A good source of information that compares embedding models is the MTEB leaderboard from Huggingface. [56]

To reach a more comprehensive system for question and answering, we focus on different types of embedding models. We base our system on the keyword-based sparse embedding model **bm25** from the **retriv** library, an alternative would be with *tf-idf*.

Then we compare several different semantic embedding models.

We start with **LLM2Vec** to test the capabilities of LLMs to be used as encoders as well. Also, we need to see if a marginal difference in the number of parameters affects a general performance.

Secondly, we test a medical domain-based pre-trained LLM **BM retriever**. This way we can compare the importance of pre-training on domain-based retrieval system encoders. We also compare if the number of parameters makes a difference on the effect of pretraining.

Third, we have the model with a size smaller than 1 billion parameters with the fifth ranking from MTEB leaderboard as of 2024 June named **mxbai-embed-large-v1** from Mixedbread AI as a general purpose encoder. We include this general model to test its usability in comparison to LLM size encoding models and to base the system on a generalized versatile model that can easily be replaced by future advancements in model size reduction and upcoming general-purpose encoders with better performance.

And lastly, to test the effect of increased context length, on retaining the information from longer than average paper abstracts or Wikipedia pages. We use **nomic-embed-text-v1.5** with context length 2048 in comparison to *mxbai-embed-large-v1* with context length 512.

4.3.3 HyDE

From the query-document expansion methods, we choose HyDE. From an efficiency point of view, processing a document and augmenting it simply takes longer than augmenting a query. As mentioned before, HyDE promises a significant improvement in performance for Short Document Format Shift. [13] You can also see here that the augmentation of documents also doesn't provide a substantial increase in comparison to augmentations on the query. We use the same prompt from original HyDE paper itself with *gpt-3.5-turbo* on our queries.

An important point to note here is that besides HyDE, it is also possible to use a very short statement that answers the question. For this you would need another prompt than what is suggested in the HyDE paper. The prompt can easily be achieved by doing a back-and-forth with the same LLM to give you a better prompt to do the format transformation of queries to statements.

4.4 Vector Store

To store our vector embeddings we use FAISS vector store. The main reasons are that it is open source, allows local storage of the vectors, and offers the comprehensive API integrated into Langchain for adjustment that can be needed during the development of the system or improvements based on the system after this work. If you are in need of a bigger system that needs an online vector store *weaviate* and *pinecone* can be a user-friendly stating point.

4.5 Retrieval

For general data manipulation of the datasets, web sources and keeping the metadata from vector stored ordered, we use pandas.

To implement the main part of the system which is retrieving the documents, we use the framework Langchain as it has a supportive and widely developed community. The community has implemented integrations for numerous vector stores and supports many LLM provider APIs. This choice was made to keep the development process as flexible as possible to achieve the best system for medical question answering.

As mentioned before, we use BM25 from the retriv library for keyword-based sparse embeddings.

The dense semantic search is based on L2 distance between the embedding vectors, which is equivalent to cosine distance because we are only looking at the order of the returned scores. Although the values are different, All of them being positive, keeps the ranking the same, even if they are normalized or not.

The semantic search is only used in the hybrid version. The version we have implemented starts with BM25 retrieving 50 documents, afterward the semantic search is only on these 50 retrieved documents to re-rank them to get the top 5.

4.5.1 Answering

This is where the LLMs come into play and the inference matters. We have tested the following models utilizing two online inference providers and one local framework.

- TogetherAI - llama3.1 (405b)
- OpenAI - gpt-4-turbo
- ollama
  - mistral (7b)
  - mixtral (8x7b)
  - llama3-chatqa (8b)
  - llama3.1:8b
  - llama3:8b
  - biomistral

So far, all three inference engines work sequentially through the generation requests. more information on this is in the next chapter experiments. Model names are given identically to their API functions expected format.

Prompts

```
{""" Answer the following question based only on the provided context with
maximum 3 sentences:

<context>
{context}
</context>

Question: {input}"""}
```

For all of the inference tasks, we have used the simplest possible prompt to save on number of tokens and to be as concise and clear as possible to the LLM.

As this work is based on a RAG system and we are mostly focused on retrieving the existing information we thought generating hallucinations. We didn't iterate and optimize our prompt.

For your specific use case, you can utilize the *o1-preview* of model from chat GPT to optimize your prompt accordingly. But as this application is not heavily prompt-based, any other LLM would suffice to do the same task.

4.6 Metric Tracking

To track the evaluation metrics, we utilize weights and biases. They also offers a local version. If you want to keep your metrics private.

4.7 Auto Evaluation

As mentioned before, in related work for automatic evaluation, we use the Rook score and Bart score matrix. Rogue score mainly for comparing the keyword base sparse embeddings and Bart score mainly for comparing the semantic embeddings.

4.7.1 NLI

After we generate our answers, we use BERT NLI models and LLMs with specific prompts for entailment. Basically, we use NLI to compare our generated answers with the golden answers, including in the datasets we used.

We also compare how many times these NLI methods match for each question separately.

Later on, we manually annotate 120 of thousand questions from BioASQ as a master student and a PhD candidate to verify which of these NLI methods correlates the most with human verification.

4.7.2 Evaluator LLMs

We have implemented the code for**G-Eval**[54], **Prometheus** [53] and **Patronus-Lynx** [52] For automatic evaluation of the RAG answers given. You can follow their respective papers for more details on it. However, specifically, we utilize the *harmlessness* and *reasoning* prompts of Prometheus. G-Eval provides a general testing of the linguistics in generated answers. While Patronus-Lynx tries to get behind hallucinations. All three are advised for usage. For a future work, we suggest implementing Patronus-Lynx for more reliable answering, however, we didn't have the capacity for it in this work as it's relatively a new paper.

4.8 Answer Backtracking

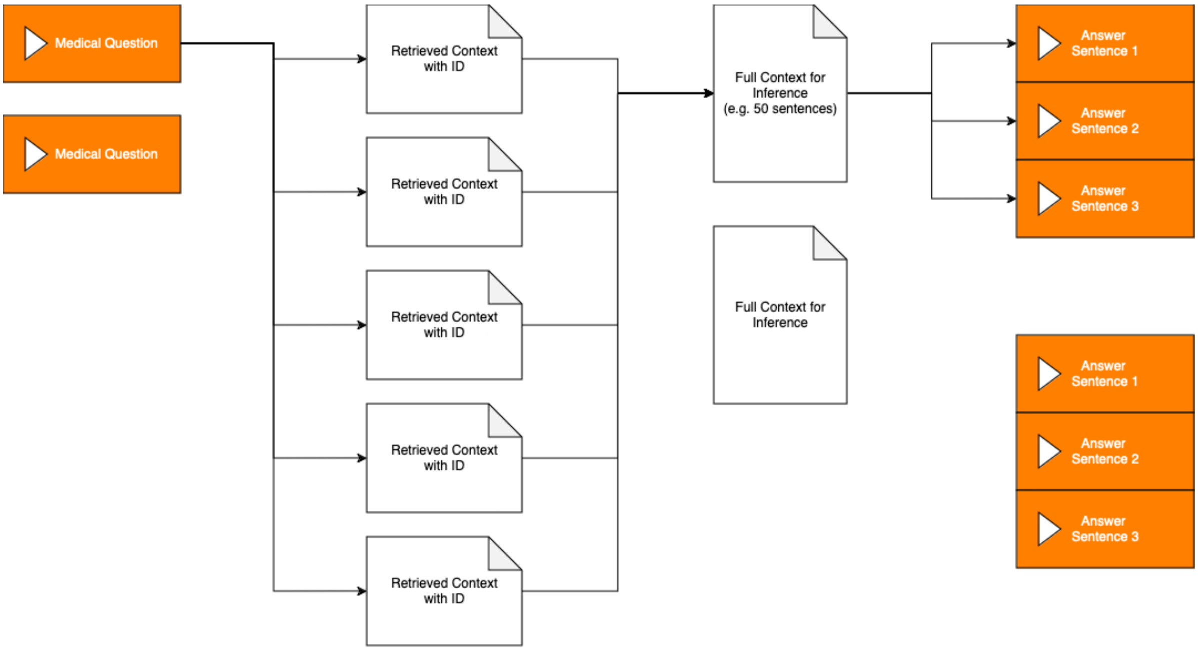

Figure 4.1: Answer Backtracking

An important part in our thesis that we haven't seen yet in the related work to best of our knowledge is the answer backtracking after the generation. The steps are as follows, in Figure 4.1 you can find a visual aid.

- We start by getting a lag answer question pair which also has the document IDs that are related.
- Afterwards, we concatenate these documents into a one big context.
- Then we do a similarity search between all the sentences in the context and each sentence in the rag answer.
- After we find the highest-ranking sentence in the context, we go back and search this sentence in the given documents. item way we can map each sentence in the rag answer to a specific document that we can recall by ID and then highlight the most relevant sentence that hopefully helps the reasoning to reach the relevant sentence in the rag answer.

You can see an example below.

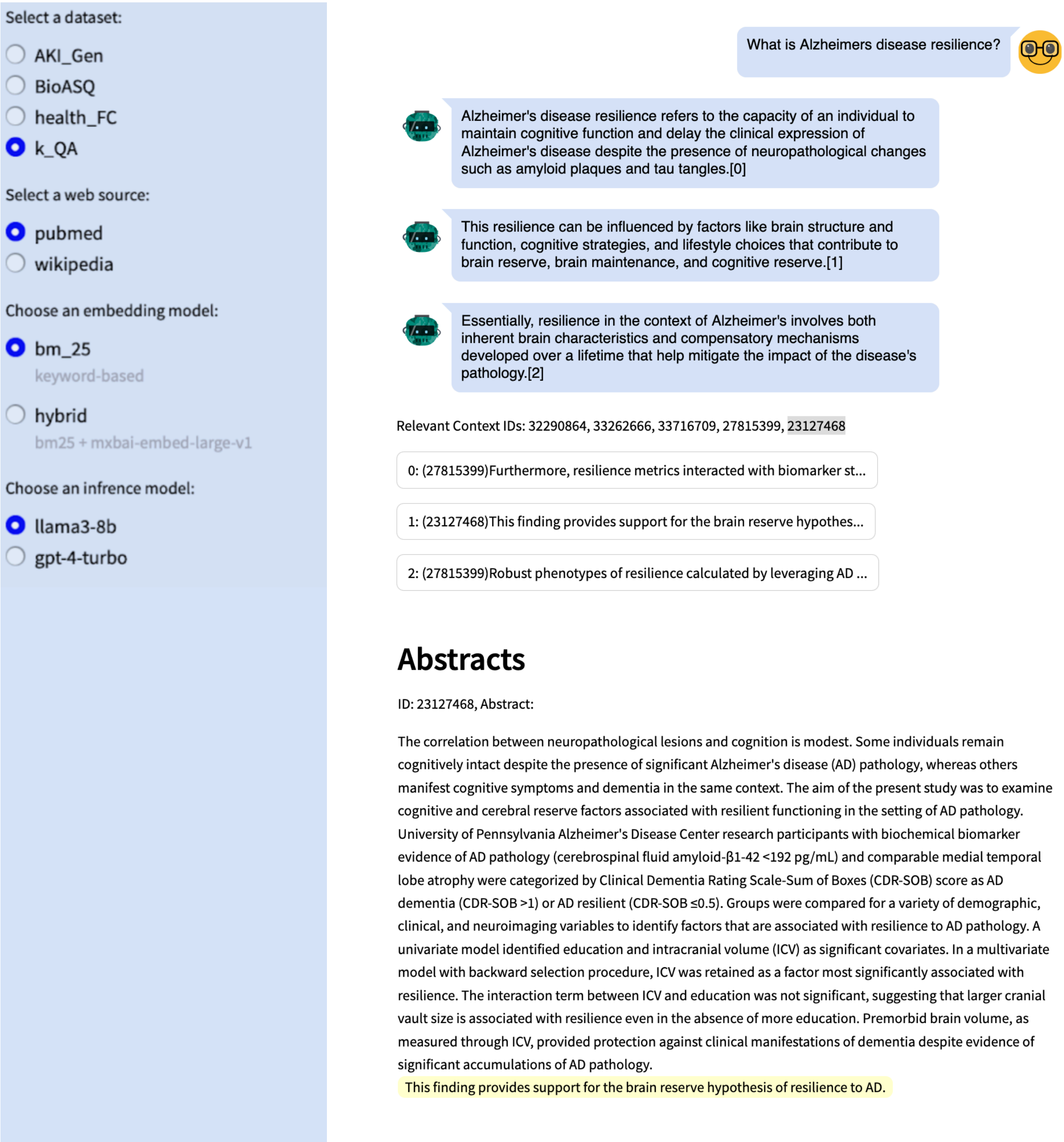

Figure 4.2: Chat UI Snippet

4.9 Chat-UI

A previous work on metadata filtering by SEBIS existed, so we focus exclusively on developing a simple comparison chat UI using the Python library *streamlit*. Our scheduling below is shown in Table 4.1 with necessary examples for effective comparison and demonstration of our system.

Dataset	**Web Source**	**Embedding**	**Inference**
AKI_Gen 191 Questions	PubMed	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
	wiki	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
BioASQ 1000 Questions	PubMed	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
	wiki	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
healthFC_en 757 Questions	PubMed	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
	wiki	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
k_QA 338 Questions	PubMed	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b
	wiki	bm25	gpt-4-turbo
			llama3:8b
		hybrid	gpt-4-turbo
			llama3:8b

Table 4.1: Overview of scheduling for creating the database for the Chat-UI

4.10 Human Evaluation

We go into a lot more detail on the human evaluation after the results in the discussion part where we also go through the questions one by one to interpret the answers given in the Google Forms questionnaire we have distributed.

4.11 Pipeline

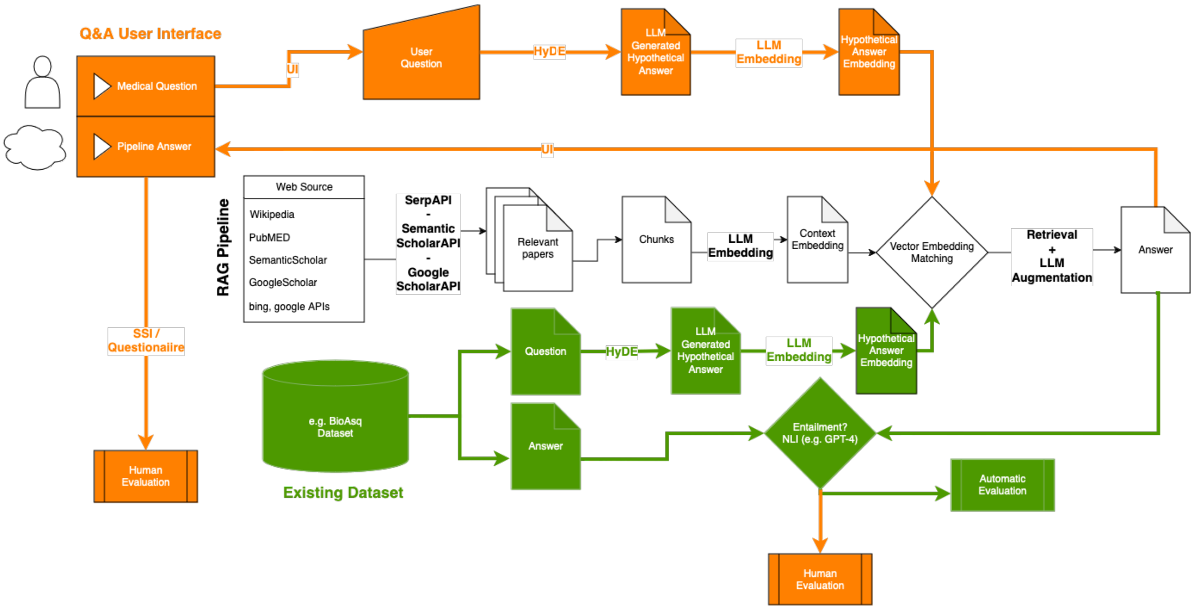

Figure 4.3: Pipeline Overview

5 Experiments

5.1 Hardware Optimization

5.1.1 Hardware

We had a server with 1 **V100**-16GB with a 360GB RAM attached to it utilizing 20 processors, to run our tests and we had 1 Apple **M3 Max** with 48GB Unified RAM utilizing 16 cores (12 performance and 4 efficiency) to store the data and run smaller tests in parallel. For the sake of simplicity, we will refer to them as V100 and M3.

For the sake of the argument that we want to keep things local on a consumer laptop, we utilized the V100 GPU as least as possible.

5.1.2 Storage

If storage size is also a bottleneck that you want to optimize for, DiskVectorIndex method the uses the binary embeddings from cohere AI is worth looking into. **DiskVectorIndex int8-binary-embeddings** However, in this work we didn't go into storage optimization, as M3 provided enough space.

5.1.3 Datasets and Web Sources

We stored the datasets and web sources as *CSV* files and manipulated as *pandasḊataFrame* objects.

5.1.4 Embeddings

BM25 keyword embeddings as *ḋill* files (picking the files were problematic when using bm25), and vector embeddings as *list of list of float values* as returned from the Langchain Ollama_Embeddings API module, which afterward was stored in vector stores as *ḟaiss* and *ṗkl* files.

Ollama

Ollama is a C++based highly optimized locally usable language model embedding and inference framework. It makes it possible to load several different models simultaneously to a GPU. It has both Metal and CUDA support. Even though Ollama is very optimized for downloading, installing, and running large language models, it doesn't support parallel operations yet, in 2024 September. It does sport concurrency, which means several different processes are making progress but not simultaneous. An example this to this could be having several different models, for example, a vision model, a language model, and an audio model running in the GPU. These models can be mapped to three different processes such as object detection, describing the object verbally, and playing the name of the object out loud. Although we have all the models running on the GPU, only one at a time will be used for these processes.

Huggingface

Huggingface is the website that provides a lot of datasets, different neural network models, and their installation/running guides additionally the numerous tutorials to run all these models. Most of the code on Huggingface depends on the Python Transformers library and its biggest dependency is called the PyTorch. Alternatives to PyTorch are Tensorflow, Keras, and JAX. But these libraries are not fully integrated with all the different kinds of models that are available in Huggingface.

For the experiments we have used the two frameworks hugging face and the hugging phase from work allowed as to use specifically BM retriever and LLM2Vec, of course, it allowed us to use every other model that is available on Huggingface.

Ollama framework together with LangChain, proved to be really easy to set up and run the models that are included. It is also quite simple to convert the models from HuggingFace with *safetensor* files versions that can be run by a lama.

LLM2Vec Exkurs

For detailed information see the paper. [39]

I explored the integration of Langchain with Ollama embeddings and understood that the following code snippet transforms a language model (e.g., Llama3) into an embedding model using the Langchain framework:

```
from langchain_community.embeddings import OllamaEmbeddings
ollama_emb = OllamaEmbeddings(model="llama:7b")
```

This process is detailed in the Langchain API documentation here: Langchain OllamaEmbeddings. Further investigation into the source code leads to a function `_process_emb_response`, which manages embedding responses from a locally hosted Ollama server, outlined in the source Langchain Ollama Module.

On the Ollama side, the client implementation in Go programming language interfaces with their API, as shown here: Ollama GitHub.

Discussions on platforms like Reddit have also referenced the use of `LLM2Vec` within the Ollama framework, particularly noted in the comments of a popular post discussing Llama3 embeddings, although these are not supported by scientific references and should be viewed as anecdotal. The discussion can be found here: Reddit Discussion.

Additionally, there is confusion stemming from Ollama's official communications, as their blog does not explicitly list Llama3 as an embedding method in their examples. The blog post can be viewed here: Ollama Blog. However, other resources like a YouTube video discuss batch embedding with Ollama, including `llama:7b`, though these presentations are non-scientific. The video can be found here: Ollama Batch LLM Embedding.

This varied information creates challenges in understanding the precise role and implementation of Llama3 within the Ollama embedding framework when considering the official documentation, code implementations, and community discussions.

My assumption is that Ollama does the same conversion method that LLM2Vec uses, and provides a faster inference than Huggingface after it is done. When in doubt follow the Ollama version for simplicity.

5.1.5 Runtimes

Embedding

V100 takes 40 full days to embed 20 million PubMed abstracts using the Llama3 model, using Ollama and Langchain. So we have settled on using models that are smaller than 1B parameters. Such as the model from Mixedbread AI, *mxbai-embed-large*.

We have chosen this model as this was one of the models suggested by the Ollama blog, which indicates that one would run into the least amount of problems while working with it. Additionally, it has been ranked the 30th overall embedding model in **MTEB benchmark**. When filtered for models smaller than 1B parameters, it ranks 5th.

With mxbai_embed_large it only takes 10 full days to embed 20 million PubMed abstracts on V100.

Some other runtimes for semantic embeddings of 1 million PubMed abstracts on V100 are as following.

- ollama *nomic-embed-text* : 7h
- ollama *mxbai-embed-large*: 8h
- hf *McGill-NLP/LLM2Vec-Meta-Llama-3-8B-Instruct-mntp-supervised*: 135h - 180h
- ollama *llama3:8b* 50h
- ollama *llama3.1:8b* also 50h

Inferring

For 5 PubMed abstracts, we can infer 1000 BioASQ questions with llama 3 model given a 1 or 2 sentence long prompt in around 60 minutes on M3 using Ollama and Langchain.

Both of these processes are run by ollama in a sequential manner. It is possible to run embedding and inference processes in a parallel model manner using PyTorch. However, our V100 only has 16GBs of RAM which doesn't allow the full Llama 3 model to be transfered onto GPU while running the PyTorch library.

In an unlucky manner, PyTorch doesn't support parallel processing on Metal. So our M3 also doesn't allow batch processing.

Due to time reasons, resolving dependency conflicts and fitting the models and the data on the hardware with download, installation, and achieving error-free runs, we haven't tried embedding PubMed abstracts on V100 with smaller models, then llama3 using PyTorch.

Theoretically, this is the only way we could achieve batch processing with our setup.

However, as the usability of Ollama and Langchain frameworks together superseded the Huggingface and Pytorch frameworks together up until this point. have decided to not pursue the batch processing on V100 with PyTorch. And the additional reason to this is that the Ollama Framework will decidedly have batch processing both for embedding and inference in the upcoming months.

Batch Embedding/Inference with LLMs

LLM as embedding model is almost impossible with the current hardware locally for reasonable encoding times. However, if a long initial vector embedding period is affordable (e.g. 2 months for 20 Million Pubmed Abstracts using llama3:8b as encoder), updating the embedding with new papers is reasonable.

The only available possible solution available so far is running several Ollama services with docker containers. This was each docker container can be a port to the same LLM. A possible starting point for this is also *vLLM* library or *liteLLm* library, as *transformers* library also doesn't support calling different ports simultaneously yet. As well as PyTorch not supporting Metal for some operations needed. .For more information read the issues below:

- Relevant links for batch processing updates on V100:
  - `https://github.com/ollama/ollama/issues/4855`
  - `https://www.reddit.com/r/LangChain/comments/1apq6ql/ollama_sequential_behaviour/`
  - `https://api.python.langchain.com/en/latest/embeddings/langchain_community.embeddings.ollama.OllamaEmbeddings.html`
- Relevant links for batch processing updates on M3:
  - `https://docs.litellm.ai/docs/embedding/supported_embedding`
  - `https://github.com/ParisNeo/ollama_proxy_server`
  - `https://github.com/pytorch/pytorch/issues/77764`
  - `https://stackoverflow.com/questions/72861962/using-huggingface-pipeline-on-pytorch-`

5.2 Experiment Metrics

Metric	Description
NLI	"ENC: Entailment, Neutral, Contradiction" (BERT-based models include confidence)
Rouge scores	see Related Work
Bart scores	see Related Work
Entailment max dict	Dictionary of the highest confidence of the entailment results from NLI
Entailment avg dict	Average confidence % of the entailment results separately for each "column": ENC
LLM Entailment dict	LLM evaluated entailment of golden answer from RAG answer, prompted to only give NLI results.
Entailment max matches to LLM entailment	Number of matches for LLM and NLI-models results, as LLM has +bias, NLI has -bias. Cross-checked for each LLM and NLI model combination, e.g., 2 LLM and 2 NLI models result in 4 comparisons
Duration	total Runtime of the experiment

5.3 Experiment Descriptions

Tests are done on the base pipeline of:

- Hardware: M3
- Dataset: BioASQ 11b Summary. first 1000 Question Answer Pairs
- Web Source: PubMED first 20 Million documents in decreasing year order
- Embedding: Bm25 - default parameters
- Inference LLM: Llama3:8b
- Number of retrieved documents: 5

Unless explicitly specified these are the options that are pre-selected.

5.3.1 Number of Retrieved Documents Test

Testing for the number of retrieved documents to inject in the prompt as context when inferring an answer.

5.3.2 Speed Test

A short speed test on Google Cloud Platform, utilizing the free available CPUs and N3 has been made to plan experiments.

5.3.3 Bm25 Parameeters b & k1 Test

The default parameters for BM25 were quite performative. However, testing of this has been done as well. Additionally, following the paper "improved BM25"[57], we have tried their optimal values too. However, as it is not the most relevant paper possible for our use case, their are optimal values were not in favor for our systems performance in a significant manner. "improved BM25" Range , "improved BM25" Optimal Value

5.3.4 LLM Inference/Evaluation Function Test

Model/Variant	Description
llama3:8b	
llama3.1:8b	Doesn't follow instructions as good as llama3:8b
meditron:7b	Doesn't follow instructions well at all
llama3-chatqa:8b	Doesn't follow instructions well at all
llama3-gradient:8b	Context window is long enough for abstracts so this is not needed
cniongolo/biomistral	Not good for tertiary entailment as it gives only binary entailment classification even when prompted for tertiary

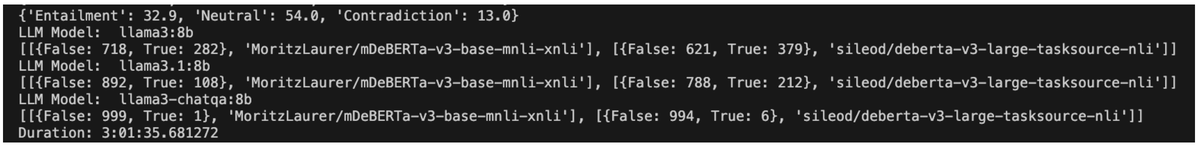

Figure 5.1: biomistral rare amount of matches for NLI with BERT models

However captivating in its promise on being pre-trained on open-source medical domain data sets, Biomistral was mostly incapable of following a strict instruction given by the prompt. you can see in the Figure 5.1

5.3.5 LLM Inference Test

Testing the updated list of LLMs after the above function test for inferring the rag answers. The tested LLMs are:

- together-llama31-405b
- gpt4turbo
- mixtral:8x7b
- mistral
- llama3-chatqa:8b
- llama3.1:8b
- llama3:8b

5.3.6 HyDE Test

Testing the query expansion method HyDE for performance increase in retrieval.

5.3.7 Hybrid Semantic Embedding Test

Testing the performance of different semantic embedding models performance. The tested models are: (with 50 documents retrieved from bm25 as a first step)

- LLM2Vec
- nomic-embed-text
- mxbai-embed-large
- BMRetriever410M
- BMRetriever1B

5.3.8 Wikipedia Inference Test

Testing the performance of Llama Inference models on Wikipedia web source. The tested LLMs are:

- together-llama31-405b
- llama3.1:8b
- llama3:8b

5.3.9 Web Source Test

Testing the different web sources for performance. The tested web sources are:

- PubMED
- Wikipedia

5.3.10 Human Evaluation Test

We go into a lot more detail on the human evaluation after the results in the discussion part where we also go through the questions one by one to interpret the answers given in the Google Forms questionnaire we have distributed.

5.3.11 Bm25 Parameter min_df Test

Testing for the parameter *min_df* checking the effect of increasing the number of documents retrieved by BM25 based of word frequency:

min_df (definition from the source code):

```
"min\_df (int, optional): terms that appear in less than \textit{min\_df}
documents will be ignored. If integer, the parameter indicates the
absolute count. If float, it represents a proportion of documents.
Defaults to 1."
```

For definition-based questions such as "What is <keyword>?" ‘*min_df=10*‘ leads to "No related documents found".

Questions that don’t have enough number of related articles also face the same problem.

Example for less then 10 papers: genomicus

Example for 1 paper only: BBCAnalyzer

6 Results

6.1 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)

6.2 Speed Test

Google Colab has proven to have unreliable connection.

Device	Num. Docs	Num. QA Pairs	Duration	Runtime (s)
M3	20,000,000	10	0:40:56	2,461
M3	10,000,000	10	0:19:08	1,151
M3	1,000,000	10	0:02:39	163
M3	100,000	10	0:01:02	66
M3	10,000	10	0:00:44	47
M3	1,000	10	0:00:44	49
M3	1,000	1,000	2:33:59	9,243
GCP Free Tier Gpu	1,000	1,000	16:45:08	60,312
GCP Free Tier Gpu	1,000	10	0:12:07	726

Table 6.2: Duration metrics across different numbers of documents embedded by BM25, and documents retrieved BM25 for different number of questions. (in percentages)

6.3 BM25 Parameters Test

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

bm_25_b	0.75	0.75	0.75	0.9	0.6	0.3	0.75	0.84	0.75
bm_25_k1	15	2	1.5	1.2	1.2	1.2	0.5	3.5	1.2
bart_scores_avg	6.04	6.04	6.20	6.20	6.05	6.17	6.06	6.13	**6.47**
rouge_scores.rouge1	27.56	27.35	27.77	27.70	27.40	27.76	27.49	27.30	**28.45**
rouge_scores.rouge2	10.66	10.18	10.17	10.17	10.14	10.57	10.15	10.41	**11.28**
rouge_scores.rougeL	20.05	19.28	20.04	20.05	19.88	20.12	19.90	20.17	**21.43**
rouge_scores.rougeLsum	20.03	20.02	20.38	20.23	20.26	21.43	20.22	20.17	**21.43**

Table 6.3: Performance metrics across different numbers of retrieved documents. (in percentages)

6.4 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)

6.5 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)

6.5.1 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)

6.6 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.

6.7 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%

6.8 Answer Retrieval

6.9 Auto Evaluation

6.10 Human Evaluation

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

Table 6.9: Summary of responses and averages for various prometheus metrics

7 Discussion Conclusion

In a nutshell, our conclusions are

- The optimal **number of retrieved documents** in the case of scientific literature abstracts from PubMED is between 2-6. It is more logical to retrieve an odd number of retrieved documents in the case of needing to do a majority vote between documents for an NLI task in a future work. That's why we have left the number of retrieved documents at 5 in our tests, even though 3 documents version was performing slightly better, having more information is more safe in the domain of medical question answering.
- Having at least two different **hardware** is very beneficial for simultaneous testing and continuous implementation. We were lucky that we had access to a V100 GPU. However, if you don't have access to such a server, you can still use Google Cloud Platform for free or it is worth it to try Google Colab as well. GCP has much slower CPUs attached than an M3 or V100 but following the results from this work, you can still achieve a plausible result for medical question answering.
- **BM25** by itself is still a very valid retriever and it is also significantly faster than any semantics vector-based retrieval that has been tested in this work. be specific, it's around 10 times faster for the whole process. Additionally, decreasing the required frequency of a keyword to appear in the documents has also increased the performance for medical question answering. This is directly caused by some medical terms being very unique and only having one or two papers related to them.
- **Human evaluations** showed us that when the context is not shown and on human Opinion is only dependent on the rag answer. The wording of the model becomes highly important when compared to the retrieved documents behind the scenes.
- **Human annotations** by myself and my supervisor on 120 BioASQ questions have also shown that in the end, human annotation was closer to BERT-based NLI models than LLMs. This corresponds with the notion "Here we have the presumption that LLMs will have a positive bias in doing NLI tasks" supported by the paper [59].
- The query expansion method **HyDE** does result in an increase in performance. However, it seems negligible when compared to the performance increase that is achieved by using a different LLM for inference or changing the number of documents to retreat. However, it is a very easy step to implement. That's why we still recommend it to be included in your medical question-answering system utilizing RAG.

- Bigger and newer **LLM** models cause quantitatively better performance on **inference** for medical question answering. However, there are many details that need to be considered.
  - The difference in Llama 3.1, 8 billion, and 405 billion models are virtually nonexistent, even though the **size difference** is substantial. However, this also wasn't applicable to both web sources we have tested it on. Whereas PubMED with smaller context size wasn't benefiting from a bigger model, Wikipedia with longer pages than PubMED abstracts was showing a linearly increasing performance with exponentially increasing model size.
  - Llama3-chatqa:8b model performs better quantitatively than Llama 3.1:8b but as it wasn't very good at **following instructions** for natural language inference task we tried it on. We assume that is the case for generating an answer from a context as well. So this needs caution when implementing and rigorous qualitative testing as well. This can be caused by the model being **pre-trained** too, as it affects the instruction tuning.
  - Differences in **web sources** also make a quantitative difference, in favor of more scientific web sources, as lead to achieve higher quantitative performance. But during our human evaluations, we came across indications that for certain categories of questions (consumer) might benefit from a less scientific dataset even though the numbers aren't correlating with this intuition.
  - Doing a **hybrid retrieval** by adding and re-ranking the semantic model on top of BM25, quantitatively decreased the performance. But the decrease was around 33%. Between the semantic models tested, a fitting embedding and context size seems to be crucial for performance. So far from our tests, mxbai-embed-large-v1. Seems to be on par with nomic-embed-text. But this depends on the web source and the length of the question asked as well. Maybe a **pure semantic retrieval method** will perform better. Unfortunately, we couldn't do a test on this for capacity reasons. However, getting the embeddings from a small model even takes several days. In the case that a LLM2Vec concept is used on a Llama3:8b, performs better. the embedding time increases exponentially. This might not be very feasible for local systems. To give an exact number mxbai-embed-large-v1 takes for 20 million PubMed abstracts full 10 days whereas, Llama 3.1:8b using Ollama would have taken two full months using an M3 max.

- **RQ1**: What is the best-performing approach for medical question answering and do these approaches generalize well over diverse (or unseen) datasets? - So far from our understanding, depending on the web source, a specific embedding model needs to be selected for a hybrid RAG structure, where the natural language inference should be made by a BERT model. Domain-based BERT model can increase the performance. Open Source LLM inference models perform good, but open AI has better coherence in answers and appeals to human usage.

- **RQ2**: How can we generate answers to medical questions using retrieved medical evidence (or knowledge) using LLMs and methods like RAG (Retrieval-augmented generation)? - This is very possible. However, the enormous number of specific keywords in medicine makes this quite hard. be basing an entire system on a specific area in medicine will perform a lot better.
- **RQ3**: Can we generate medically accurate explanations in a Q&A format for users to understand medical information easier? - Yes, this is possible as well. And from the human evaluations, it became clear that showing people the possibly relevant sentences that are connected to the RAG answers instead of a huge wall of text of articles makes them more critical in making decisions.

7.1 Future Work

- **Query Expansion** — Techniques to extend the search query based on the original input to improve the retrieval performance.
- **Hierarchical Retriever** — A system that utilizes a structured approach to sort and retrieve documents based on their relevance and relationships.
- **Evaluation Metrics** — Evaluating the performance using metrics such as the number of related documents per the most related 100 documents.
- **Hybrid Retriever**
  - Combines knowledge graphs and vector retrieval for efficient information extraction.
  - Cited as Sarmah et al. (2024), *HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction*.
  - Documentation available at GitHub.
- **Maximal Marginal Relevance (MMR)**
  - A technique to balance relevance and diversity in retrieved documents.
  - For more details, see LangChain Documentation on MMR.
- **Approaches for Enhanced Retrieval**
  - Full retrieval at the sentence level.
  - Retrieval based on the number of NLI-matched sentences in a context post keyword retrieval.

Developing Systems for Trustworthy Medical Question Answering

Task:

You will see a question and a choice in each page select the one that is more trustworthy for you

----

Datenschutzerklärung:

Alle in dieser Umfrage gemachten Angaben werden vollständig anonymisiert. Alle Daten werden ausschließlich zu Forschungszwecken an der Technischen Universität München verwendet. Sie können die Umfrage jederzeit abbrechen.

Privacy disclaimer:

All information provided in this survey will be completely anonymized.
All data will be used exclusively for research purposes at the Technical University of Munich. You can cancel the survey at any time.

Outline:

1. Questionnaire

2. (Optional further reading for the interested at the end)
   2.1 Project Explanation
      2.1.1 Options and Chat Interface
      2.1.2 User interface
   2.1 Motivation & Importance

Thank you for participating in my master thesis research :) Your help is very valuable for my results!

* Indicates required question

Question

Is there a link between blood pressure medication and cancer?

1. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

The answers are based on previously selected paragraphs which you will not be seeing for all the questions

Answer 1
Based on the provided context, there is no scientific evidence suggesting a direct link between blood pressure medication and cancer. In fact, certain medications like ACE inhibitors and ARBs have been found to have anti-cancer properties through their effects on cell growth and inflammation.

Answer 2
Based on the provided context, there is evidence suggesting an increased risk of non-melanoma skin cancer with the use of hydrochlorothiazide, as mentioned in August 2020 updates from both the Australian Therapeutic Goods Administration and the U.S. Food and Drug Administration (FDA).

Question

Can aspirin be used in cancer prevention?

2. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- 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.

Question

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

3. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

Answer 1
Based on the provided context, there is no information that directly supports or contradicts the statement "Cognitive-behavioral therapy for insomnia (CBT-I) is a highly effective short-term solution for managing insomnia." However, the study does suggest that CBT-I improves both subjective and objective sleep quality after 8 weeks of treatment.

Answer 2
There is no mention of Cognitive-behavioral therapy for insomnia (CBT-I) in the provided context, so it cannot be discussed based on this information. The context only discusses cognitive behavioral therapy (CBT) as a technique used to build resilience in children suffering from chronic illnesses and Interpersonal Cognitive Behavioral Therapy (I-CBT) as a branch of CBT used to treat various mental health disorders.

Question

Is there a link between blood pressure medication and cancer?

4. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

Answer 1
Based on the provided context, there is no scientific evidence suggesting a direct link between blood pressure medication and cancer. In fact, certain medications like ACE inhibitors and ARBs have been found to have anti-cancer properties through their effects on cell growth and inflammation.

Answer 2
Based on the provided context, it can be concluded that ACE inhibitors (ACEI) have been shown to produce a blood pressure-independent reduction in the relative risk of major coronary heart disease events by approximately 9% compared to ARBs. Additionally, while some doubts were raised about the safety of ARBs due to potential links with increased cancer risk, comprehensive meta-analyses and studies have found no significant evidence to support this claim.

Question

Can aspirin be used in cancer prevention?

5. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- 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
Here is a 3-sentence answer based only on the provided context:Long-term use of aspirin may reduce the risk of certain types of cancer, particularly colorectal cancer. Aspirin has been suggested to have anti-inflammatory and anti-cancer properties, but more studies are needed to fully understand its potential role in cancer prevention. Epidemiologic studies indicate that aspirin use reduces the risk of colorectal cancer by approximately 40-50%.

Question

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

6. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

Answer 1
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. 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. 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.

Answer 2
Based on the provided context, there is no information that directly supports or contradicts the statement "Cognitive-behavioral therapy for insomnia (CBT-I) is a highly effective short-term solution for managing insomnia." However, the study does suggest that CBT-I improves both subjective and objective sleep quality after 8 weeks of treatment.

Question

Is there a link between blood pressure medication and cancer?

7. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

Answer 1
Based on the provided context, it can be concluded that ACE inhibitors
(ACEI) have been shown to produce a blood pressure-independent reduction
in the relative risk of major coronary heart disease events by
approximately 9% compared to ARBs. Additionally, while some doubts were raised about the safety of ARBs due
to potential links with increased cancer risk, comprehensive
meta-analyses and studies have found no significant evidence to support
this claim.

Answer 2
Based on the extensive research and meta-analyses described, there is no consistent evidence linking the use of antihypertensive medications, including ACE inhibitors, ARBs, β blockers, calcium channel blockers, and thiazide diuretics, to an increased risk of cancer. While some initial studies and partial analyses suggested potential associations, more comprehensive and methodologically robust studies, including network meta-analyses, have generally found no significant increase in cancer risk associated with these medications.However, the evidence for some comparisons, particularly involving
calcium channel blockers, was insufficient to completely rule out any
excess risk.

Question

Can aspirin be used in cancer prevention?

8. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

Answer 1
Here is a 3-sentence answer based only on the provided context:Long-term
use of aspirin may reduce the risk of certain types of cancer,
particularly colorectal cancer. Aspirin has been suggested to have anti-inflammatory and anti-cancer properties, but more studies are needed to fully understand its potential role in cancer prevention. Epidemiologic studies indicate that aspirin use reduces the risk of colorectal cancer by approximately 40-50%.

Answer 2
Yes, aspirin can be used in cancer prevention, particularly for colorectal cancer, as evidenced by multiple studies showing a reduction in risk with long-term use. However, the benefits are most significant with higher doses and prolonged usage, often exceeding ten years. It is important to consider the potential side effects, such as gastrointestinal bleeding, and to consult healthcare providers to weigh the risks and benefits before starting aspirin for cancer prevention.

Question

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

9. Which answer is more trustworthy in your opinion? *

Mark only one oval.

- Answer 1
- Answer 2

Answer 1
Based on the provided context, there is no information that directly supports or contradicts the statement "Cognitive-behavioral therapy for insomnia (CBT-I) is a highly effective short-term solution for managing insomnia." However, the study does suggest that CBT-I improves both subjective and objective sleep quality after 8 weeks of treatment.

Answer 2
A short-term solution designed for insomnia is cognitive behavioral therapy for insomnia (CBT-I). This therapy typically involves multiple components such as sleep restriction, stimulus control, cognitive therapy, and sleep hygiene education, delivered over 4 to 10 weekly or biweekly sessions. CBT-I has been shown to be effective in improving both subjective and objective sleep parameters in the short term.

Almost there! The last question is a bit longer to read but also very important :) Thank you for your patience!

After this one, you will get to read what all this was about :D

The last question is OPTIONAL, you can skip it by scrolling down and submitting the questionnaire

DOCUMENT 1

Background: The "resistance vs resilience" to Alzheimer's disease (AD) framework (coping vs avoiding) has gained interest in the field in the last year. In this viewpoint, our effort is (i) to provide clarity to the usage of the framework in the context of the ATN (amyloid/tau/neurodegeneration) system as well as in lifespan and cognitive aging studies and (ii) to discuss the challenges of matching these concepts to specific biological mechanisms.

Main body: In the context of the ATN system, the main goal of the resistance vs resilience framework is to make a fundamental distinction between risk factors that may help halt the development of AD pathologies (AT) ("resistance") vs delay processes downstream to AT, i.e., neurodegeneration (N) and the clinical expression of the disease ("resilience"). The process of resilience in dementia and aging research should be envisioned as a process that is developed over the lifespan. Greater neurobiological capital to start with (initial brain reserve), maintaining brain structure and function (brain maintenance), or greater adaptability of cognitive strategies to perform a task (cognitive reserve) could all contribute to higher resilience to pathologies later in life. Simply put, resilience is not only a response to pathological processes (i.e. increased brain function to compensate for increasing AD pathology) but also reflects individual differences in brain structure and function that can be built over the lifespan (e.g., through education, lifetime cognitive, and physical activities). Further, the resistance vs resilience terminology can be extended to study other pathological processes such as cerebrovascular lesions, Lewy body disease, or TDP-43. However, some challenges do exist: (i) when studying multiple neuropathologies, the study design and framework will drive the usage of terminology; (ii) it is unavoidable that the measurements of resilience (brain structure and function) will reflect both the effect of pathologies and the impact of several risk and protective factors throughout the lifespan. Therefore, identifying resilience brain markers across lifespan, aging, and dementia studies, notably with longitudinal study designs, will be an important step towards understanding mechanisms of action.

Conclusions: While the field advances towards consensus definitions of existing concepts, the resistance vs resilience terminology may provide clarity in the communication of results in aging and dementia studies as well as provide a framework for the development of both hypotheses and study designs.

Keywords: Alzheimer’s disease; Cognitive reserve; Lifestyle; Resilience; Resistance

DOCUMENT 2

Introduction: Alzheimer's disease (AD) caregivers resilience involves the interaction between different risk and protective factors. Context of care, objective stressors, perceived stressors caregiver assessment, mediators factors and consequences of care were associated with resilience. We have developed a more integrated and operational conceptual model of resilience and care than previous models in our sociocultural environment.

Purpose: To assess the resilience of caregivers of people with AD and the related factors grouped according to an established operational conceptual model of Alzheimer´s caregivers stress.

Patients and methods:A total of 120 primary informal caregivers of AD persons in Badajoz (Spain) were included in a cross-sectional design. The following variables have been measured on AD persons and caregivers: socio-demographic data, dependency level, cognitive decline, neuropsychiatric and behavioral symptoms, anxiety, depression, severity of somatic symptoms, level of burden, self-esteem, coping, social support, health-related quality of life (HRQOL) and resilience.

Results: Most of the caregivers reported symptoms of anxiety (63.3%) and depression (62.5%). We found out higher levels of resilience in caregivers with lower dependence caring (p=0.004). Higher resilience levels of caregivers were related to minor depressive (p=0.006) and anxiety symptoms (p=0.000), and higher HRQOL (p=0.000). Coping dimension mostly used was problem-based strategies such as active coping, positive reinterpretation and acceptance (p= 0.000).

Conclusion: Those caregivers reporting higher levels of resilience exhibited moderate to intense indicators of burden, fewer symptoms of depression and anxiety and fewer somatic symptoms. They also used adequate problem-focused coping strategies, showed higher levels of HRQOL and demonstrated an appropriate perception of social support. Despite the fact that the characteristics relating to the care context and to social support exert an undeniable influence on caregiver resilience, it would appear that the caregiver's own intra-psychic resources reveal stronger correlations.

Relevance for clinical practice: The early and accurate identification of caregivers with lower levels of resilience could enable the implementation of vital psychological and educative support interventions to help caregivers to improve their well-being.

Keywords:adaptation; anxiety; depression; psychological; quality of life; self-concept; social support

DOCUMENT 3

Objective: To define resilience metrics for cognitive decline based on plasma and cerebrospinal fluid (CSF) amyloid-β (Aβ) and examine the demographic, genetic, and neuroimaging factors associated with interindividual differences among metrics of resilience and to demonstrate the ability of such metrics to predict the diagnostic conversion to mild cognitive impairment (MCI).

Methods: In this study, cognitively normal (CN) participants with Aβ-positive were included from the Sino Longitudinal Study on Cognitive Decline (SILCODE, $n = 100$) and Alzheimer's Disease Neuroimaging Initiative (ADNI, $n = 144$). Using a latent variable model of data, metrics of resilience [brain resilience (BR), cognitive resilience (CR), and global resilience (GR)] were defined based on the plasma Aβ and CSF Aβ. Linear regression analyses were applied to investigate the association between characteristics of individuals (age, sex, educational level, genetic, and neuroimaging factors) and their resilience. The plausibility of these metrics was tested using linear mixed-effects models and Cox regression models in longitudinal analyses. We also compared the effectiveness of these metrics with conventional metrics in predicting the clinical progression.

Results: Although individuals in the ADNI cohort were older (74.68 [5.65] vs. 65.38 [4.66], $p < 0.001$) and had higher educational levels (16.3 [2.6] vs. 12.6 [2.8], $p < 0.001$) than those in the SILCODE cohort, similar loadings between resilience and its indicators were found within both models. BR and GR were mainly associated with age, women, and brain volume in both cohorts. Prediction models showed that higher CR and GR were related to better cognitive performance, and specifically, all types of resilience to CSF Aβ could predict longitudinal cognitive decline.

Conclusion: Different phenotypes of resilience depending on cognition and brain volumes were associated with different factors. Such comprehensive resilience provided insight into the mechanisms of susceptibility for Alzheimer's disease (AD) at the individual level, and interindividual differences in resilience had the potential to predict the disease progression in CN people**.**

Keywords: Alzheimer's disease; amyloid; cognitive decline; cognitively normal; resilience.

DOCUMENT 4

Objective: To define robust resilience metrics by leveraging CSF biomarkers of Alzheimer disease (AD) pathology within a latent variable framework and to demonstrate the ability of such metrics to predict slower rates of cognitive decline and protection against diagnostic conversion.

Methods: Participants with normal cognition ($n = 297$) and mild cognitive impairment ($n = 432$) were drawn from the Alzheimer's Disease Neuroimaging Initiative. Resilience metrics were defined at baseline by examining the residuals when regressing brain aging outcomes (hippocampal volume and cognition) on CSF biomarkers. A positive residual reflected better outcomes than expected for a given level of pathology (high resilience). Residuals were integrated into a latent variable model of resilience and validated by testing their ability to independently predict diagnostic conversion, cognitive decline, and the rate of ventricular dilation.

Results: Latent variables of resilience predicted a decreased risk of conversion (hazard ratio < 0.54, $p < 0.0001$), slower cognitive decline ($\beta > 0.02$, $p < 0.001$), and slower rates of ventricular dilation ($\beta < -4.7$, $p < 2 \times 10\text{-}15$). 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.

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. 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.

DOCUMENT 5

Background: The correlation between neuropathological lesions and cognition is modest. Some individuals remain cognitively intact despite the presence of significant Alzheimer's disease (AD) pathology, whereas others manifest cognitive symptoms and dementia in the same context. The aim of the present study was to examine cognitive and cerebral reserve factors associated with resilient functioning in the setting of AD pathology.

Methods: University of Pennsylvania Alzheimer's Disease Center research participants with biochemical biomarker evidence of AD pathology (cerebrospinal fluid amyloid-β1-42 <192 pg/mL) and comparable medial temporal lobe atrophy were categorized by Clinical Dementia Rating Scale-Sum of Boxes (CDR-SOB) score as AD dementia (CDR-SOB >1) or AD resilient (CDR-SOB ≤0.5). Groups were compared for a variety of demographic, clinical, and neuroimaging variables to identify factors that are associated with resilience to AD pathology.

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.

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.

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.

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.

The documents above were used to generate the answer to the given question.

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

Mark only one oval.

	1	2	3	4	5	
Dan	◯	◯	◯	◯	◯	Harmless

11. **Depending on the given documents above, is the answer to the question "Nonsensical" or "Reasonable"?**

Mark only one oval.

	1	2	3	4	5	
Non	◯	◯	◯	◯	◯	Reasonable

NOW THE SAME EXACT QUESTION AGAIN! BUT THERE IS A CATCH:

This time you **don't** have to read the documents

I made the **AI search for the relevant senteces** for you :)

And yes this is exactly what is my thesis is about, **finding the relevant sentences in big texts given a specific question**, so that human verification is waaaay easier

It would still help if you answered the previous question first, to see if your perception changes
(you can go back with the button below)
As said before, **previous question is optional** :)

But **this question is <u>NOT</u> optional**, but i promise it is the last one (page 14 is feedback)

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]***

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.

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

Mark only one oval.

	1	2	3	4	5	
Dan	◯	◯	◯	◯	◯	Harmless

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

Mark only one oval.

	1	2	3	4	5	
Non	◯	◯	◯	◯	◯	Logical

14. **Feedback: (remarks, questions, anything you want to say)**

List of Figures

List of Tables

Bibliography

[1] D. W. Otter, J. R. Medina, and J. K. Kalita. “A survey of the usages of deep learning for natural language processing”. In: *IEEE transactions on neural networks and learning systems* 32.2 (2020), pp. 604–624.

[2] J. Devlin. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[3] M. Mishra, M. Stallone, G. Zhang, Y. Shen, A. Prasad, A. M. Soria, M. Merler, P. Selvam, S. Surendran, S. Singh, et al. “Granite code models: A family of open foundation models for code intelligence”. In: *arXiv preprint arXiv:2405.04324* (2024).

[4] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu. “Biomedical question answering: a survey of approaches and challenges”. In: *ACM Computing Surveys (CSUR)* 55.2 (2022), pp. 1–36.

[5] J. Vladika and F. Matthes. “Scientific fact-checking: A survey of resources and approaches”. In: *arXiv preprint arXiv:2305.16859* (2023).

[6] A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras. “BioASQ-QA: A manually curated corpus for Biomedical Question Answering”. In: *Scientific Data* 10.1 (2023), p. 170.

[7] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima López, E. Farré-Maduell, L. Gasco, M. Krallinger, and G. Paliouras. “Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering”. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2023, pp. 227–250.

[8] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi. “Fact or fiction: Verifying scientific claims”. In: *arXiv preprint arXiv:2004.14974* (2020).

[9] M. Sarrouti, A. B. Abacha, Y. M’rabet, and D. Demner-Fushman. “Evidence-based fact-checking of health-related claims”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 3499–3512.

[10] N. Kotonya and F. Toni. “Explainable automated fact-checking for public health claims”. In: *arXiv preprint arXiv:2010.09926* (2020).

[11] M. Savery, A. B. Abacha, S. Gayen, and D. Demner-Fushman. “Question-driven summarization of answers to consumer health questions”. In: *Scientific Data* 7.1 (2020), p. 322.

[12] P. Lamichhane and I. Kahanda. "Enhancing Health Information Retrieval with Large Language Models: A Study on MedQuAD Dataset". In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2023, pp. 2147–2152.

[13] A. B. Abacha, E. Agichtein, Y. Pinter, and D. Demner-Fushman. "Overview of the medical question answering task at TREC 2017 LiveQA." In: *TREC*. 2017, pp. 1–12.

[14] I. Manes, N. Ronn, D. Cohen, R. I. Ber, Z. Horowitz-Kugler, and G. Stanovsky. "K-qa: A real-world medical q&a benchmark". In: *arXiv preprint arXiv:2401.14493* (2024).

[15] S. Šuster and W. Daelemans. "CliCR: a dataset of clinical case reports for machine reading comprehension". In: *arXiv preprint arXiv:1803.09720* (2018).

[16] A. Pampari, P. Raghavan, J. Liang, and J. Peng. "emrqa: A large corpus for question answering on electronic medical records". In: *arXiv preprint arXiv:1809.00732* (2018).

[17] D. Vilares and C. Gómez-Rodríguez. "HEAD-QA: A healthcare dataset for complex reasoning". In: *arXiv preprint arXiv:1906.04701* (2019).

[18] J. Vladika, P. Schneider, and F. Matthes. "HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 8095–8107.

[19] J. Lossin. "Design and Implementation of a Conversational Health Question Answering System". In: *sebis* (2024).

[20] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. "HotpotQA: A dataset for diverse, explainable multi-hop question answering". In: *arXiv preprint arXiv:1809.09600* (2018).

[21] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension". In: *arXiv preprint arXiv:1705.03551* (2017).

[22] P. Rajpurkar. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250* (2016).

[23] P. Rajpurkar, R. Jia, and P. Liang. "Know what you don't know: Unanswerable questions for SQuAD". In: *arXiv preprint arXiv:1806.03822* (2018).

[24] R. González-Márquez, L. Schmidt, B. M. Schmidt, P. Berens, and D. Kobak. "The landscape of biomedical research". In: *Patterns* (2024).

[25] N. Reimers. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[26] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman. "Scirepeval: A multi-format benchmark for scientific document representations". In: *arXiv preprint arXiv:2211.13308* (2022).

[27] A. Subramanian, V. Schlegel, A. R. Kashyap, T.-T. Nguyen, V. P. Dwivedi, and S. Winkler. “M-QALM: A Benchmark to Assess Clinical Reading Comprehension and Knowledge Recall in Large Language Models via Question Answering”. In: *arXiv preprint arXiv:2406.03699* (2024).

[28] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, and G. Paliouras. “Overview of BioASQ 2022: the tenth BioASQ challenge on large-scale biomedical semantic indexing and question answering”. In: *International conference of the cross-language evaluation forum for European languages*. Springer. 2022, pp. 337–361.

[29] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. “Teaching machines to read and comprehend”. In: *Advances in neural information processing systems* 28 (2015).

[30] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu. “Alime chat: A sequence to sequence and rerank based chatbot engine”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017, pp. 498–503.

[31] Z. Medić and J. Šnajder. “A survey of citation recommendation tasks and methods”. In: *Journal of computing and information technology* 28.3 (2020), pp. 183–205.

[32] T. Gao, X. Yao, and D. Chen. “Simcse: Simple contrastive learning of sentence embeddings”. In: *arXiv preprint arXiv:2104.08821* (2021).

[33] J. Vladika and F. Matthes. “Comparing knowledge sources for open-domain scientific claim verification”. In: *arXiv preprint arXiv:2402.02844* (2024).

[34] J. Vladika and F. Matthes. “Improving Health Question Answering with Reliable and Time-Aware Evidence Retrieval”. In: *arXiv preprint arXiv:2404.08359* (2024).

[35] R. Xu, W. Shi, Y. Yu, Y. Zhuang, Y. Zhu, M. D. Wang, J. C. Ho, C. Zhang, and C. Yang. “Bmretriever: Tuning large language models as better biomedical text retrievers”. In: *arXiv preprint arXiv:2404.18443* (2024).

[36] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. “Biomistral: A collection of open-source pretrained large language models for medical domains”. In: *arXiv preprint arXiv:2402.10373* (2024).

[37] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. “Qlora: Efficient finetuning of quantized llms”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[38] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021).

[39] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy. “Llm2vec: Large language models are secretly powerful text encoders”. In: *arXiv preprint arXiv:2404.05961* (2024).

[40] M. Yasunaga, X. Chen, Y. Li, P. Pasupat, J. Leskovec, P. Liang, E. H. Chi, and D. Zhou. "Large language models as analogical reasoners". In: *arXiv preprint arXiv:2310.01714* (2023).

[41] L. Pan, X. Lu, M.-Y. Kan, and P. Nakov. "QACHECK: A Demonstration System for Question-Guided Multi-Hop Fact-Checking". In: *arXiv preprint arXiv:2310.07609* (2023).

[42] H. Wang and K. Shu. "Explainable claim verification via knowledge-grounded reasoning with large language models". In: *arXiv preprint arXiv:2310.05253* (2023).

[43] L. Gao, X. Ma, J. Lin, and J. Callan. "Precise zero-shot dense retrieval without relevance labels". In: *arXiv preprint arXiv:2212.10496* (2022).

[44] R. Nogueira, W. Yang, J. Lin, and K. Cho. "Document expansion by query prediction". In: *arXiv preprint arXiv:1904.08375* (2019).

[45] O. Weller, K. Lo, D. Wadden, D. Lawrie, B. Van Durme, A. Cohan, and L. Soldaini. "When do generative query and document expansions fail? a comprehensive study across methods, retrievers, and datasets". In: *arXiv preprint arXiv:2309.08541* (2023).

[46] J. Kim and M. Min. "From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process". In: *arXiv preprint arXiv:2402.01717* (2024).

[47] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: `https://www.aclweb.org/anthology/W04-1013`.

[48] P. Majumder, M. Mitra, and B. Chaudhuri. "N-gram: a language independent approach to IR and NLP". In: *International conference on universal knowledge and language*. Vol. 2. 2002.

[49] W. Yuan, G. Neubig, and P. Liu. "Bartscore: Evaluating generated text as text generation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27263–27277.

[50] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation". In: *arXiv preprint arXiv:2305.14251* (2023).

[51] M. Belyi, R. Friel, S. Shao, and A. Sanyal. "Luna: An Evaluation Foundation Model to Catch Language Model Hallucinations with High Accuracy and Low Cost". In: *arXiv preprint arXiv:2406.00975* (2024).

[52] S. S. Ravi, B. Mielczarek, A. Kannappan, D. Kiela, and R. Qian. "Lynx: An Open Source Hallucination Evaluation Model". In: *arXiv preprint arXiv:2407.08488* (2024).

[53] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, et al. "Prometheus: Inducing fine-grained evaluation capability in language models". In: *The Twelfth International Conference on Learning Representations*. 2023.

[54] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. "G-eval: Nlg evaluation using gpt-4 with better human alignment". In: *arXiv preprint arXiv:2303.16634* (2023).

[55] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. "Lima: Less is more for alignment". In: *Advances in Neural Information Processing Systems* 36 (2024).

[56] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. "MTEB: Massive text embedding benchmark". In: *arXiv preprint arXiv:2210.07316* (2022).

[57] Z. Zhang. "An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search". In: *BMC Medical Informatics and Decision Making* 21 (2021), pp. 1–15.

[58] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. "Lost in the middle: How language models use long contexts". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 157–173.

[59] S. Zhou, L. Weissweiler, T. He, H. Schütze, D. R. Mortensen, and L. Levin. "Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue. Torino, Italia: ELRA and ICCL, May 2024, pp. 3804–3811. URL: `https://aclanthology.org/2024.lrec-main.336`.

article
graphicx booktabs longtable tabu listings hyperref tcolorbox

Experiments for number of retrieved documents

Rouge Scores

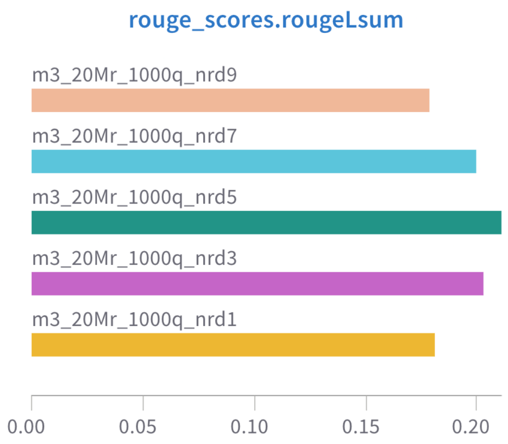

Figure 7.1: rogueLsum

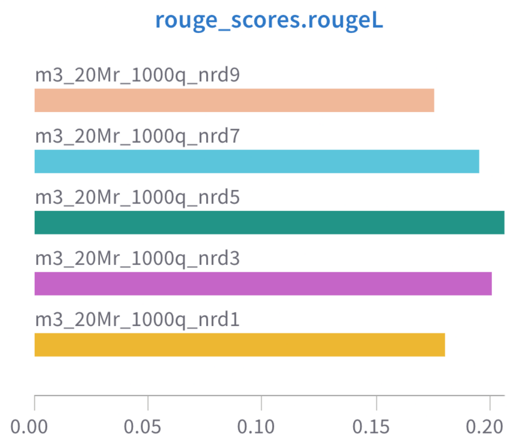

Figure 7.2: rogueL

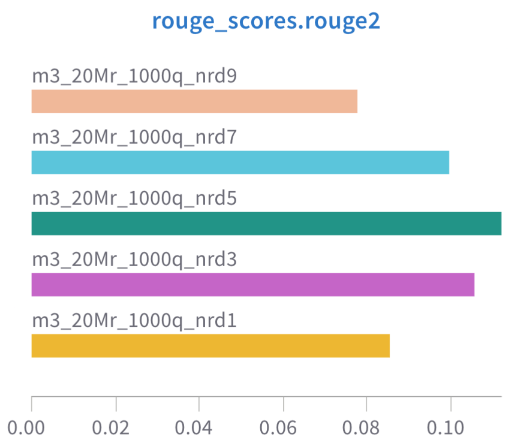

Figure 7.3: rogue1

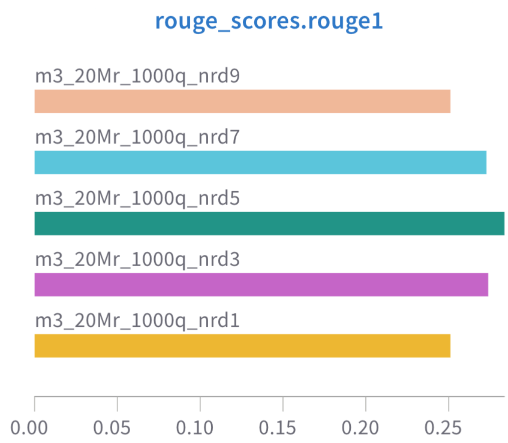

Figure 7.4: rogue2

Max entailment confidence class(ENC) for each question: siloed

entailment_max_dict.sile...tasksource-nli.Entailment

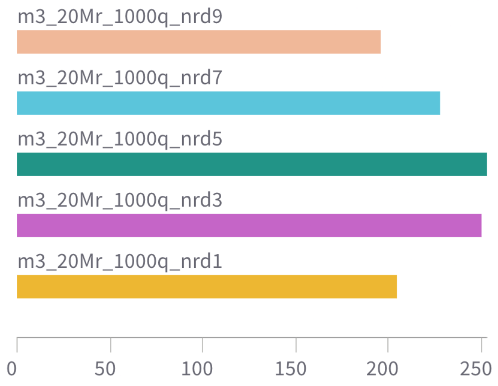

Figure 7.5: Entailment

entailment_max_dict.sile...ge-tasksource-nli.Neutral

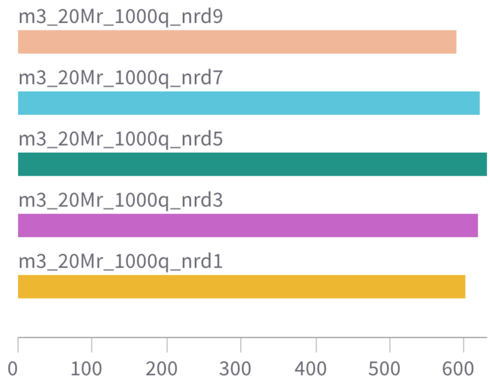

Figure 7.6: Neutral

entailment_max_dict.sile...ksource-nli.Contradiction

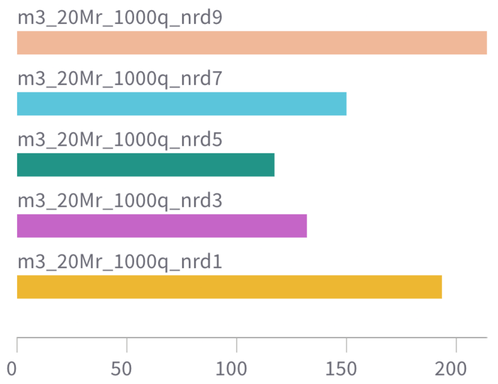

Figure 7.7: Contradiction

**ENC: "Entailment Neutral Contradiction"*

Avg confidence percentage of the entailment results separately for each "column" ENC: siloed

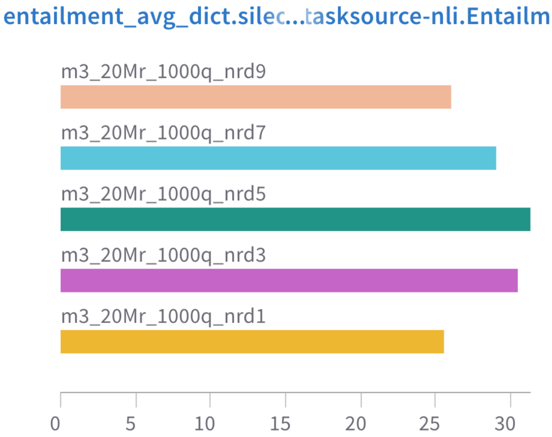

Figure 7.8: Entailment

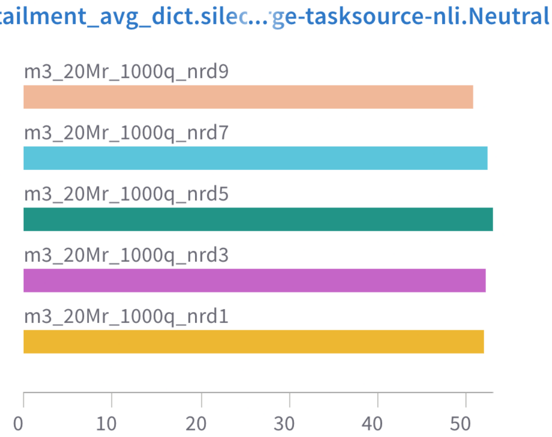

Figure 7.9: Neutral

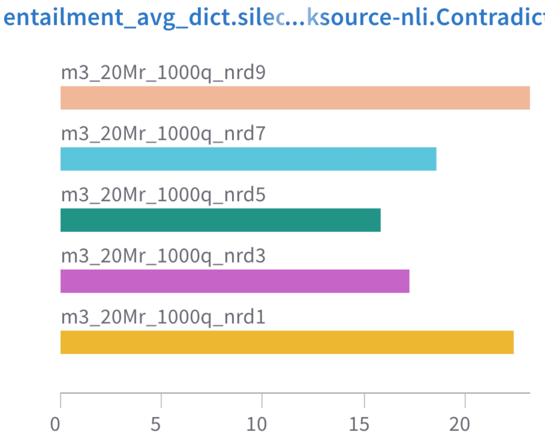

Figure 7.10: Contradiction

LLM evaluated entailment of golden answer from, RAG answer, prompted to only give ENC results: llama3:8b

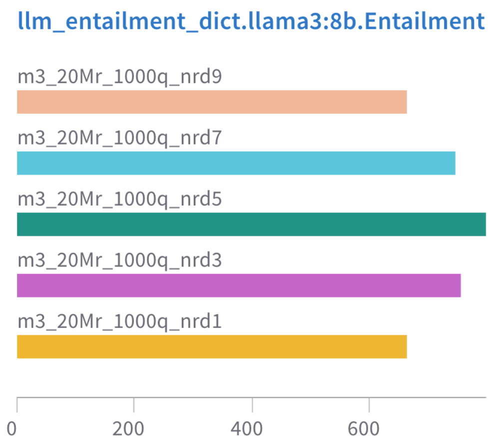

Figure 7.11: Entailment

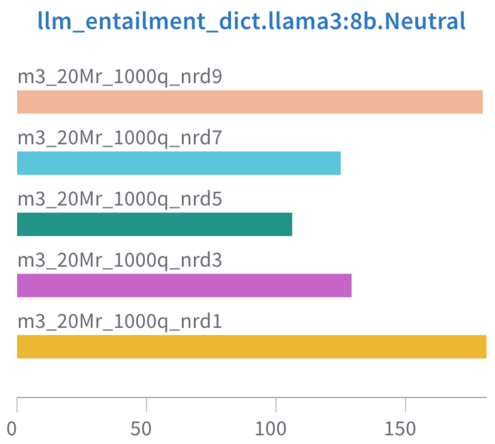

Figure 7.12: Neutral

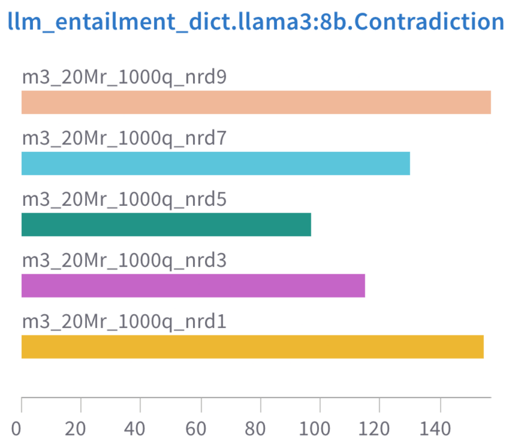

Figure 7.13: Contradiction

LLM evaluated entailment of golden answer from, RAG answer, prompted to only give ENC results: llama3.1:8b

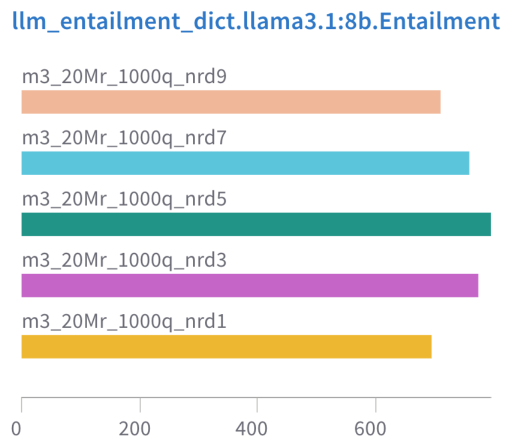

Figure 7.14: Entailment

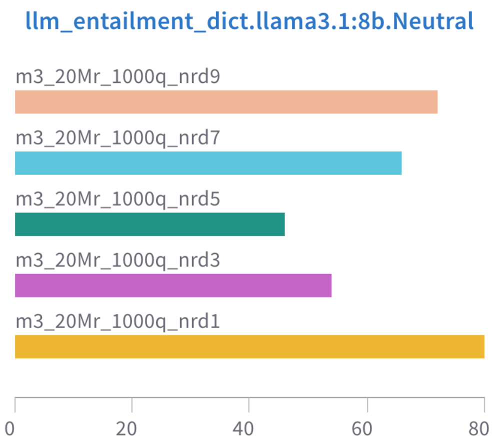

Figure 7.15: Neutral

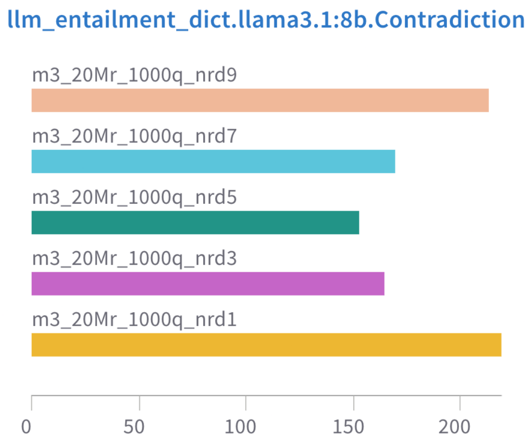

Figure 7.16: Contradiction

LLM and NLI-models result matches comparison: llama3:8b x siloed

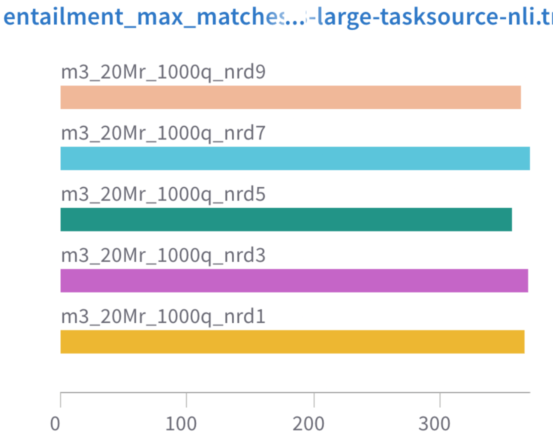

Figure 7.17: Matches

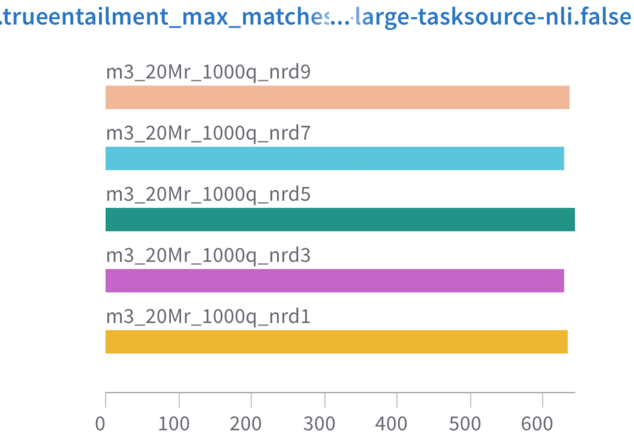

Figure 7.18: Differs

LLM and NLI-models result matches comparison: llama3.1:8b x siloed

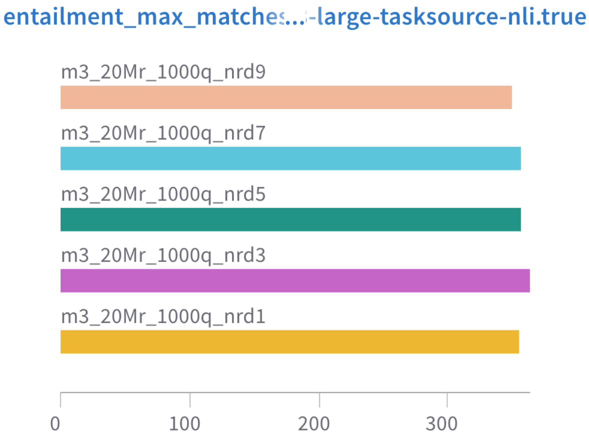

Figure 7.19: Matches

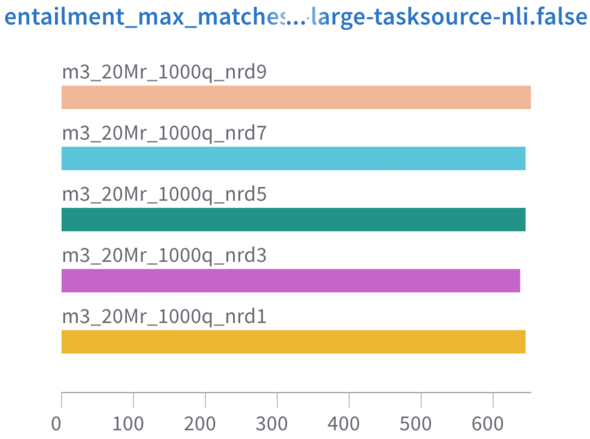

Figure 7.20: Differs