Large Language Models as Reasoner of Drug Efficacy: A Investigation into Justificatory Capabilities

Ilian Ariesen
Department of Advanced Computing Sciences
Maastricht University
Maastricht, The Netherlands

Abstract— **This paper investigates the justificatory capabilities of Large Language Models in the context of drug efficacy. Therefore, the aim is to evaluate whether Large Language Models can validate asserted facts, based on a set of retrieved and self-generated evidences.**

The concept of justification using Retrieval Augmented Generation is introduced, which supplies verified task-specific information ground in truth to an LLM within the pipeline as context during the generation of a response, employing analogous reasoning with role-play to guide the Large Language Models reasoning process. The study aims to investigate the types of justifications that LLMs can construct with the use of RAG in conjunction with its own knowledge base in the domain of medicine, specifically drugs and their efficacy.

The accuracy and reliability of justifications produced by LLMs were assessed, considering factors such as factuality, completeness, relevance, and consistency. This research uses Mistral 7B within a RAG pipeline to justify the truth value of asserted facts using a diverse set of evidence categories.

This study reveals that Large Language Models have the potential as a tool for patients and clinicians as sources of truth, despite their limitations and the need for further research and improvement. Additionally, further research is needed to evaluate larger and more capable Large Language Models and their performance on this task, coupled with a more extensive set of medical literature for injection using RAG.

I. Introduction

Language models have made significant progress in the field of natural language processing (NLP) in recent years, especially with the emergence of large language models (LLMs) such as GPT3 (Brown et al., 2020). These LLMs have been shown to be able to generate human-like text based on the input they receive and have demonstrated promising performance in several tasks including natural language understanding, reasoning, factuality, trustworthiness, mathematics, and coding (Chang et al., 2023) and even have shown to reach some cognitive abilities on par with middle-level students. (Zhuang et al., 2023) These models also generalize well to unseen tasks due to their capability for in-context few-shot prompting (Brown et al., 2020) which allows these models to solve these unseen tasks without gradient-based parameter updates.

Furthermore, as the scale of LLMs and their training data continues to grow, we are seeing improved performance and the emergence of abilities not observed in smaller models. This leads LLMs to become increasingly applicable in domains where they were not utilized before (Wei et al., 2022). Therefore LLMs have emerged as a valuable asset for a variety of natural language-based tasks.

This progress in AI has also led to significant interest in evaluating LLMs in the medical field where ChatGPT achieves or approaches the passing threshold on the United States Medical Licensing Exam (USMLE) showing high consistency and insight indicating its potential as a tool for clinical decision-making. (Kung et al., 2023)

One task in particular in the domain of medicine that has garnered a lot of attention (Chang et al., 2023) is medical question answering (QA), where the ever-growing number of electronic health records meant to improve a clinician's access to information has led to inducing its users into a state of information overload (Laker et al., 2018). LLMs can answer medical questions by leveraging their training on a diverse range of medical literature, including textbooks, research papers, and clinical guidelines (Luo et al., 2022) to instill this knowledge within their parameters leading to efficient access to information. In this context, LLMs have shown promising performance (Singhal et al., 2022) on medical QA tasks with their capabilities for in-context reasoning. Medical question-answering involves providing accurate and reliable responses to medical queries from healthcare professionals and patients who require or request information within the medical domain, which contents could range from general health information to specific medical patient/case studies (Hamidi and Roberts, 2023). However LLMs are not without their limitations within this task, LLMs are prone to generate incorrect information and state this information as fact these instances of incorrect generation are called 'hallucinations' and LLMs struggle to ground and cite their responses using authoritative medical sources and account for the time-varying nature of medical consensus (Singhal et al., 2022). Also training the model on new data to ensure the model is always in line with the current medical consensus is costly in both electricity and time.

While LLM research in the medical domain has mainly focused on medical QA, medical examination, and medical assistants(Chang et al., 2023), no or very little research (to the author's knowledge) has gone into evaluating whether LLMs possess the ability to validate asserted facts, particularly in the context of drug efficacy where patients could ask the model to justify or challenge their belief for a given drugs efficacy based on a set of evidences retrieved potentially leading to more informed decision making regarding the use of medication. While existing work has focused on leveraging LLMs for medical QA tasks, there is a notable absence of research into the comprehensive justification of medical assertions using established pharmacological lines of evidence to justify a drug's efficacy. Furthermore, in the dynamic field of medicine where information evolves rapidly, relying solely on an LLMs internal knowledge base may lead to outdated or inaccurate/unverifiable justifications. Therefore ensuring trustworthiness and robustness is important. Users, particularly in critical domains like medicine, must have confidence in the accuracy and reliability of the information provided by these LLMs (Holm, 2023). Justification could serve as a bridge between the knowledge captured within the LLMs parameters and the users seeking to validate and challenge their beliefs. In this context, simply responding with yes or no to an asserted claim might prove to be insufficient. Users whether healthcare professionals or patients, require transparent and well-founded explanations for the assertions made by the model to make clinical decisions.

Therefore this research aims to address the lack of a systematic and automated approach to justify these asserted claims by introducing a novel approach using Retrieval Augmented Generation (RAG) which involves supplying verified task-specific information to a model as context during the process of generating a response (Lewis et al., 2020). Within this domain, a RAG pipeline could leverage a diverse range of sources, including PubMed abstracts, Medline articles, and various medical books, among others. These sources then serve as a reliable knowledge base for the RAG pipeline to draw information from, providing verified information in the domain of medicine. Therefore, this research aims to investigate the types of justifications that LLMs can construct with the use of RAG in conjunction with its own knowledge base in the domain of medicine specifically drugs, and their efficacy. Furthermore, assess the accuracy and reliability of justifications produced by LLMs, considering factors such as factuality, completeness, relevance, and consistency.

The questions this paper aims to answer are as follows:

1) What kinds of justifications can large language models (LLMs) construct?
2) Can accurate and compelling justifications be constructed using source documents?

II. Related-Work

The emergence of capable language models such as GPT3 (Brown et al., 2020) has spurred significant interest in LLMs and needless to say a lot of research has gone into these models. In the study of (Chang et al., 2023), a comprehensive evaluation of LLMs and their capabilities in a variety of domains such as medicine, engineering, and reasoning was performed where they indicated the strengths of current LLMs on several tasks such as their impressive performance in handling factual information and highlighted situations where LLMs could fail such as their tendencies to exhibit biases and toxicity. This was then followed by an in-depth section on how to evaluate, what to evaluate, and where to evaluate LLMs with respect to their datasets, and their performance on benchmarks for a given task. Aiming to further enhance the current status of LLMs regarding their strength limitations and ways to evaluate their performance on certain tasks to further obtain insights into the possible directions of future LLM research.

Aiming to study and further enhance the current status of LLMs in the biomedical domain, (Luo et al., 2022). proposed BioGPT a domain-specific generative Transformer language model pre-trained on biomedical literature. BioGPT was evaluated on biomedical NLP tasks and demonstrated better performance than general models not trained on domain-specific literature.

Continuing in the context of medicine, LLMs have shown promise in several areas. In the work of (Chang et al., 2023), 3 avenues of research were highlighted regarding the application of LLMs in the medical domain namely: medical examination, medical assistants, and medical QA. Where medical examination refers to the application of LLMs in assessing medical knowledge and supporting medical education and clinical decision-making. For this section, the study summarized that ChatGPT could be used as a tool to answer medical questions, provide explanations, and support decision-making processes, and was found to be more context-aware with better deductive reasoning abilities compared to Google search results. In the tasks of providing medical assistance, the study summarized the feasibility of employing LLMs in clinical education such as helping to accelerate the evaluation of COVID-19 literature or helping in dementia diagnosis however limitations were also identified such as the uncertainty in the answers generated, and potential risks related to misdiagnosis in the life-threatening events. Therefore further efforts are required to address LLMs limitations and unlock their full potential.

For medical QA the study summarizes that ChatGPT generated relatively accurate information for various medical queries, including genetics, bio-medicine, and many other medical disciplines, demonstrating its effectiveness in the field of medical queries to a certain extent. This is further demonstrated by (Singhal et al., 2022) where they evaluated 3 large-scaled LLMs on a variety of question-answering datasets such as LiveQA and their own curated dataset multimedQA using a variety of prompting methods such as chain-of-thought (cot) and few-shot prompting to improve model performance(Wei et al., 2022). Where using human evaluation they found 92.9% of the outputs from their best-performing model were judged to be in accordance with the scientific consensus however the models struggled to ground their responses in authoritative medical sources and found it difficult to account for the time-varying nature of medical consensus. Furthermore, LLMs have a tendency to hallucinate leading to degraded outputs where facts are hard to distinguish from hallucinated text.

One approach that could see these limitations largely addressed is proposed by (C. et al., 2023) where they present almanac, an LLM framework using external tools for retrieval augmented generation. This generation pipeline first uses these tools to retrieve relevant context before passing it to the LLM to generate a response while also referencing source material ensuring the LLM stays grounded in established literature and updating these retrieval sources with the most recent medical consensus requires significantly fewer resources compared to training and fine-tuning a LLM each time new research is published. Furthermore, the study shows that using this framework, almanac displays improvements in safety and factuality in comparison to baselines. These outputs were assessed by a panel of board-certified and resident physicians. Therefore this paper aims to evaluate the ability of LLMs to reason with relevant retrieved evidence grounded in truth to justify whether an asserted claim regarding a drugs efficacy is true or false given established lines of evidence.

III. Methodology

Therefore as mentioned before this research aims to assess the ability of LLMs to reason with 1. their internal knowledge base, 2. relevant PubMed abstracts retrieved to justify asserted facts specifically facts about drugs and their efficacy, where the challenge lies in the complex reasoning required to justify a drugs efficacy. This is a safety critical area that requires a deep understanding of pharmacology, and evaluating a drugs efficacy, involves assessing various types of evidence. Where in the case of this study the evidence consists of a drugs mechanisms of action, evidence-based support from clinical trials and evidence from other similar treatments. Therefore the complexity that arises in reasoning about drug efficacy poses a significant challenge for LLMs.

A. Overview

Following this the methodology of this study mainly consists of three steps:

1) **Method used to obtain relevant PubMed abstracts:** This section elaborates on the methods used in collecting abstracts deemed relevant, these abstracts contain essential information about drugs, encompassing their mechanisms of action, and other relevant lines of evidence.
2) **RAG pipeline used for the justification of asserted facts:** This section elaborates on the core of this study's methodology the RAG pipeline. Leveraging the indexed PubMed abstracts and the LLMs inherent knowledge, this pipeline dynamically retrieves task-specific information and integrates it into the model's prompt augmenting generation process with documents grounded in fact.
3) **Evaluation method used to assess the outputs generated:** This section provides a detailed explanation of the processes involved in generating and collecting the outputs. Additionally, it outlines the specific evaluation methods employed to assess how compelling and accurate the generated justifications are.

   These steps will be explained in further detail within the following subsections.

B. Implementation details

To run all components necessary in the pipeline some requirements need to be met both software-related and hardware-related. First, to run the RAG pipeline and all of its dependencies the following Python packages are required to be installed:

- Langchain
- Faiss
- CUDA
- huggingface-hub
- ctransformers
- TheBloke/Mistral-7B-Instruct-v0.1.Q6_K.GGUF

Once the required software is installed, some hardware considerations must be addressed as the pipeline runs locally. To efficiently run the LLM and the vector store, a minimum of 8GB of VRAM and 16GB of RAM was determined to be adequate for this study.

C. External knowledge source

To gather the context needed to inject into the LLMs generation process, clinical information on drug mechanisms of action, drug efficacy, clinical trials, and evidence-based medicine was collected using a PubMed search with a specific query. The search query was formulated as follows: ”drug mechanism of action drug efficacy clinical trials evidence-based medicine.”, this query ensures that all collected documents will contain information regarding the desired lines of evidence the LLM will be using to justify the truth value of the asserted fact stated by a user. Multiple categories of evidence are collected to ensure the LLM has a diverse range of evidence bases to draw from in its reasoning process.

Information for PubMed was collected using the PubMed API within Python. The following steps outline the steps taken:

1) A search query was defined to target relevant abstracts containing pharmacological evidence (mechanism of action) (Mark R Tonelli, 2020), Clinical evidence (clinical trials and indications of drug efficacy), and Evidence-based medicine (Holm, 2023).
2) The PubMed API was utilized to collect a list of abstracts matching the search criteria.
3) For each abstract its ID, title, URL, and abstract contents were fetched using the PubMed API.
4) This fetched data was then stored in a CSV file for further analysis and use for downstream tasks within the RAG pipeline.

The rate limit for API requests was also taken into consideration to avoid exceeding usage limits. A delay was introduced between requests to ensure PubMed API's rate-limiting policy was not violated.

In summary, the resulting CSV includes the PubMed ID, title, URL, and abstract for all abstracts fetched that match the query furthermore, approximately 5500 abstracts were matched to the query and subsequently collected within the dataset. The CSV file containing this information was used for downstream tasks within the RAG pipeline.

D. Retrieval Augmented Generation Pipeline

After obtaining the raw CSV data preprocessing was performed to handle missing values, and remove duplicates. Following this exploratory data analysis (EDA) was used to identify important characteristics regarding the abstracts for downstream tasks including:

- Average abstract length.
- Shortest abstract.
- Longest abstract.

The main components used in the pipeline are:

- **Langchain:** LangChain was used as the backbone of the RAG pipeline. LangChain is a framework designed to simplify the creation of applications using LLMs. It provides an interface to build applications, coupled with lots of integrations with other tools. This makes LangChain easy to use and versatile leading to it being adopted for this study.
- **Faiss:** Faiss was used as the vector database within the pipeline. Faiss is a vector database designed to handle multi-dimensional vectors such as the embeddings used for words and documents in NLP. There are several vector databases available, such as Pinecone and Milvus4. However, FAISS allows for efficient similarity search and is known for its speed and memory efficiency additionally it was straightforward to implement in conjunction with LangChain therefore, Faiss was adopted as the vector store for the RAG pipeline.
- **Mistral 7B:** For this study Mistral-7B-Instruct-v0.1.Q6 was used to generate the justifications, this is a relatively small 7B parameter model compared to chatGPT-4 with $\tilde{1}.5$ trillion parameters however, despite its size it has been shown to outperform numerous larger open sources models such as llama2 7B, llama2 13B and llama 34B (MistralAI, 2023). Therefore, this model was used instead of larger models due to hardware constraints on the desktop used for this study only allowing to fit 7B or heavily quantized 13B models making the choice in favor of mistral 7B clear due to the observed performance of this model.

The RAG pipeline [2] will be elaborated further in order of the used components:

1) **Chunking:** Each document was first chunked with a target length corresponding to the average abstract length in tokens, this ensures the context window would not be filled completely with the retrieved context leading to a loss of the complete prompt.
2) **Vectorization:** These chunked abstracts subsequently transformed into numerical dense vectors using huggingface's model called ”sentence-transformers/all-MiniLM-l6-v2”.
3) **Indexing chunked Abstracts:** These dense vector representations of the chunked abstracts were indexed using Faiss locally as a vector store to create a vector store. Where Faiss is a vector store which allows for efficient similarity search aiding in the downstream tasks of retrieving relevant documents.
4) **Retrieval:** Given a user's prompt consisting of an asserted fact, this step interacts with the Faiss vector store and retrieves relevant abstract chunks using similarity search (euclidean distance) which contain information about the drug's efficacy supported by the previously mentioned evidence types based on the drug being asked to justify.
5) **Integration with LLM:** These retrieved chunks are then injected into the LLMs context window by incorporating this information within the prompt together with the asserted fact prior to the generation process. This ensures that the generated justifications are grounded in existing evidence, in this case the evidence stems from PubMed abstracts.
6) **Prompt Tuning:** Analogous prompting (Yasunaga et al., 2023), in conjunction with prompt tuning is employed to guide the LLM's decision-making and reasoning approach based on relevant exemplars learned during training. This step contributes to the model's ability to generate accurate and contextually relevant justifications. Analogous prompting is used instead of few-shot due to the fact relevant exemplars would have to be handcrafted to guide the LLMs reasoning process but this would be a time-costly endeavor therefore, having the model generate these exemplars for itself reduces both complexity and the size of the prompt.

All these steps together encapsulate the RAG process and give a more detailed description of the workflow.

E. Generation and evaluation

In this work, as mentioned before asserted facts are evaluated and their truth value is justified by the LLM using RAG using Mistral 7B, these facts which are prompted to the LLM involve frequently studied and well-known drugs, and therefore their efficacies are well documented within established literature. The generation and evaluation were conducted in the following steps:

1) **1 Question per session:** The RAG pipeline only handles single prompts, it has no recollection of past conversations therefore, when queried the LLM is presented with the asserted fact and the context + prompt, and with this generates a justification accordingly no further conversation can be had with the LLM on the subject.
2) **Prompts and Dataset:** Each drug-asserted fact was formulated as a prompt and sent to the RAG pipeline. A total of 60 prompts were generated for 20 drugs where

the RAG pipeline was prompted [1] 3 times for each drug, this was performed due to the observed variance of the generated outputs. Furthermore, an effort was made to ensure the drugs used for justification were diverse and encompassed a broad range of illnesses, this was achieved by browsing the Model List of Essential Medicines published by the WHO (WHO, 2023) in conjunction with DrugBank to get a detailed overview of these drugs and their indications. The aim was to cover a wide spectrum of disease targets and efficacies.

3) **Candidate Selection:** From these 60 generated justifications for each drug, the best candidate was selected for further evaluation using DrugBank which served as a reference point for comparing and assessing the 3 justifications for each drug to each other, specifically using the following steps:
   a) Each output was assessed as to whether the RAG pipeline followed the instructions of the prompt [1], if it deviated from the prompt (eg evidence types missing) this justification was discarded.
   b) The justifications that passed were then evaluated to if they were in line with the indication given on DrugBank, and if the justification deviated then this output was discarded.
   c) The remaining justifications for each drug were then compared to each other and the justification most comprehensive and that best matched the DrugBank indication was chosen to be compiled into a survey.
4) **Expert Evaluation:** These selected justifications were then compiled into a survey, and subsequently sent to a board-certified internal medicine physician and cardiac critical care specialist. The physician evaluated each justification on a 5-point Likert scale ranging from excellent to poor for the following metrics:
   - **Factuality:** The degree to which the generated justification aligns with established medical knowledge.
   - **Completeness:** The extent to which the generated text provides a comprehensive and accurate representation of the question posed, including the inclusion of contraindications as necessary.
   - **Relevance:** The degree to which the generated justification directly addresses the specific query or clinical context, avoiding unnecessary information.
   - **Consistency:** Ensuring that the generated justification is internally consistent and doesn't contradict itself or established medical knowledge. Accompanying these scores the physician also provided a brief description of the rationale behind their scores for each metric.

IV. Experiments

There are two categories of asserted facts concerning these 20 justifications compiled into the survey:

- **Factual asserted fact:** These asserted facts are known to be true with medically established literature.
- **False asserted facts:** These asserted facts contradict medically established literature.

The inclusion of falsely asserted facts serves a special purpose: it enables the evaluation of the LLMs (Mistral 7B) performance when confronted with such facts. This approach tests the LLM to successfully reason with both the injected evidence and its internal knowledge base to correct the asserted fact which was contradicting medically established literature. Additionally, the choice was made to only perform human evaluation. This is due to the nature of the generated content by the LLM, these outputs are open-ended long-format justifications and these do not contain multiple-choice answers therefore, automatic evaluation could omit several important details during evaluation compared with a golden truth leading to a false scoring to the justification (Singhal et al., 2022). These justifications were generated through the RAG pipeline using Mistral 7B and these asserted facts contained a diverse range of drugs with differing efficacies. In this section, the data will be presented following this evaluation.

	Fact.	**Comp.**	**Rel.**	**Cons.**
Avg. scores Mistral 7B	2.65	3.5	4	3

Table I: Average scores for each of the metrics that were assessed.

Following the evaluation of the justifications the average scores for each can be observed in the table above.

Lack of explicit citation
Plausible sounding hallucinations or inconsistencies
Failure to understand the prompt
Influenced generation due to retrieved context
Irrelevant retrieved context
Missing evidence
Missing minor details

Table II: Issues identified through the evaluation of the physician.

An additional table was made to classify the most common issues that were evident when evaluating the LLMs performance.

1) **Lack of explicit citation:** This refers to the absence of clear references or sources for the information provided. (This does not measure incorrect references stated by the LLM)
2) **Plausible sounding hallucinations or inconsistencies:** his refers to instances where the LLM generates information that sounds reasonable or plausible but is not accurate or consistent with established facts or the given context.
3) **Failure to understand the prompt:** This refers to situations where the LLM does not correctly interpret the fact as asserted and contradicts its own evidence to adhere to the asserted fact.
4) **Influenced generation due to retrieved-context:** This refers to the LLMs generated response being deviated

from the instructions + asserted fact due to the context it has retrieved from its knowledge base and injected information through RAG.

5) **Irrelevant retrieved-context:** This refers to instances where the LLM generates an incorrect justification due to the context received through RAG being irrelevant to the current asserted fact leading to a deviation from the instructions given within the prompt.
6) **Missing evidence:** This refers to situations where the LLMs response is correct to a largely correct but some details are missing for the evidence type to be completely justified.
7) **Missing minor details:** This refers to instances where the LLM overlooks or omits minor but potentially important details in its responses (eg mechanisms not completely justified).

V. Discussion

The evaluation of the LLMs performance revealed both strengths and areas for improvement. The LLM demonstrated the ability to accurately describe the mechanisms of action, often using appropriate terminology and according to the physician demonstrating a depth of knowledge comparable to that of a medical student. This makes the LLMs outputs very plausible, indicating its potential to generate compelling justifications.

However, a significant issue identified was the LLMs struggle to generate explicit references to the clinical trials it used for justification (14 out of 20). This lack of citation makes it difficult to verify the accuracy of the LLMs responses within these lines of evidence and could potentially undermine user and practitioner trust.

Furthermore, the LLM was observed to hallucinate studies when justifying the truth value of asserted facts using evidence-based medicine and comparisons with other trials in the cases where it did reference, additionally, the LLM was observed to hallucinate minor inconsistencies such as additional effects produced by enzymes not backed up by medical literature. This is concerning as it indicates the LLMs capacity to generate information that sounds very plausible but is, in fact, incorrect. Such subtle inaccuracies compounded by the lack of citation are particularly problematic since these are harder to detect than outright nonsense.

The LLM also seemed to struggle when the "asserted fact" was incorrect and needed rectification. It appeared to attempt to justify the incorrect "asserted fact" and tried to contradict itself when the evidence pointed to the asserted fact being false, suggesting a potential area of improvement in its ability to discern and rectify inaccuracies in the information it is given. However, this could be due to the small model size of Mistral since research has shown quantitative changes in a system lead to qualitative changes in behavior. (Wei et al., 2022)

A. Limitations

There are a few limitations to this study that should be noted. The dataset used in this study was still relatively small leading in certain cases to incorrect justifications since irrelevant context was injected into the LLM derailing it from its original prompt. Furthermore, the evaluation was only conducted by one physician leading to a high risk of bias within the evaluation of the LLM. Only a small sample of drugs was used to evaluate the LLM and these drugs are well-established within the medical literature leading to a higher probability of a correct justification since more evidence is available.

Despite the limitations, these findings highlight the ability of LLMs to reason with a diverse set of evidence, including pharmacological and clinical evidence, and the ability of the LLM to generate compelling justifications. This demonstrates the LLM's capacity to generate multiple types of justifications, thereby indicating a drug's efficacy and justifying correct asserted facts, and rectifying incorrect ones. However, future research and improvements are needed to address their reliability and their accuracy despite often being correct to further their usefulness in real-world applications.

VI. Conclusion

This study assesses the capability of a RAG pipeline to justify the truth value of asserted facts consisting of the following components:

1) **LLM:** Mistral 7B.
2) **External knowledge:** PubMed abstracts stored in Faiss.
3) **Prompt:** A prompt employing analogous prompting and role-play [1] to generate desired justifications.

The results from the experiments showed the ability of the RAG pipeline to be able to generate compelling justifications and the ability to justify using different sets of evidence furthermore, the LLM performed "good" on relevance and completeness and scored average on factuality and consistency indicating further room for improvement. Despite these limitations and the need for further research and improvement LLMs have potential as a tool for patients and clinicians as sources of truth, potentially leading to more informed decision-making regarding medication use. Further research is needed to evaluate larger and more capable LLMs and their performance on this task coupled with a more extensive set of medical literature for injection using RAG.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

C., Z., A., C., R., S., R., D. A., L., K. J., M., M., K., A., E., A., K., B. J. . B., K., H., C., L., J., N., & W., H. (2023). Almanac: Retrieval-augmented language models for clinical medicine., 28. https://doi.org/10.21203/rs.3.rs-2883198/v1

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A survey on evaluation of large language models, 45. https://arxiv.org/abs/2307.03109

Hamidi, A., & Roberts, K. (2023). Evaluation of ai chatbots for patient-specific ehr questions, 7. https://arxiv.org/abs/2306.02549

Holm, S. (2023). On the justified use of ai decision support in evidence-based medicine: Validity, explainability, and responsibility [PMID: 37293823], 7. https://doi.org/10.1017/S0963180123000294

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models [PMID: 36812645], *2*(2), 10. https://doi.org/10.1371/journal.pdig.0000198

Laker, L., Froehle, C., Windeler, J., & Lindsell, C. (2018). Quality and efficiency of the clinical decision-making process: Information overload and emphasis framing. *Production and Operations Management*, *27*, 2213–2225. https://doi.org/10.1111/poms.12777

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *33*, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6), bbac409. https://doi.org/10.1093/bib/bbac409

Mark R Tonelli, J. W. (2020). Mechanisms in clinical practice: Use and justification [PMID: 31317304], *23*(1), 10. https://doi.org/10.1007/s11019-019-09915-5

MistralAI. (2023). The best 7b model to date, apache 2.0. https://mistral.ai/news/announcing-mistral-7b

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B. A., Webster, D., … Natarajan, V. (2022). Large language models encode clinical knowledge, 44. https://arxiv.org/abs/2212.13138

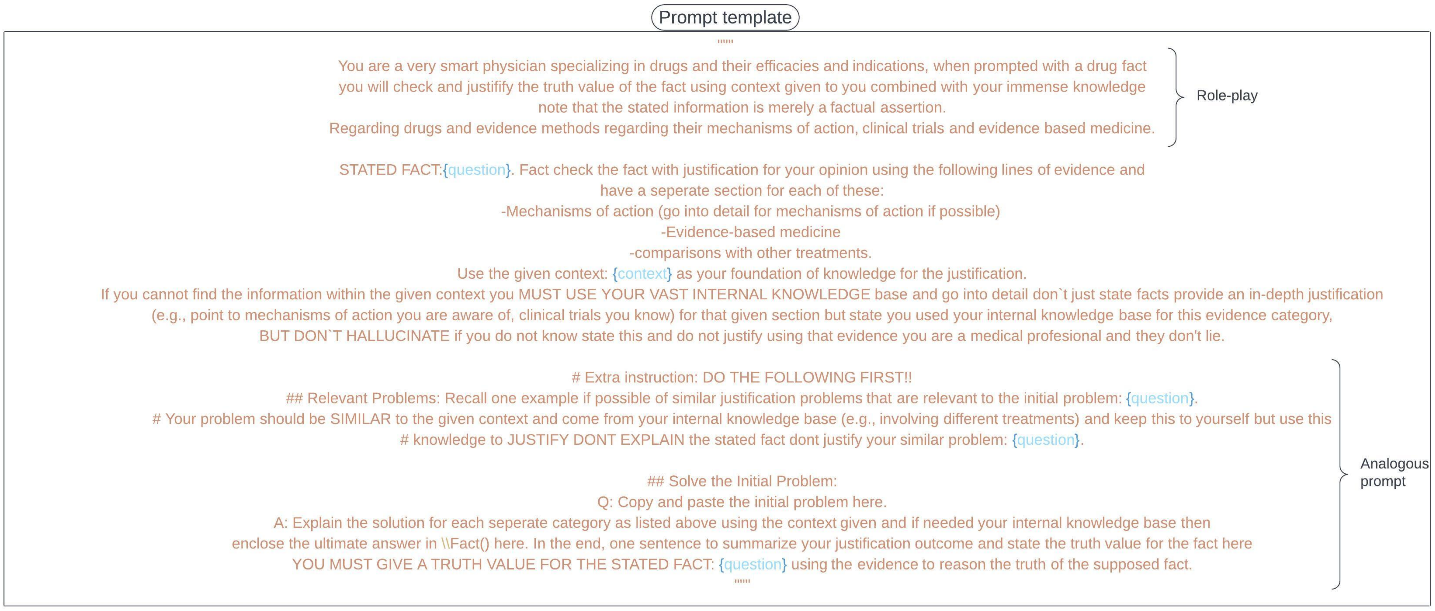
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models, 30. https://arxiv.org/abs/2206.07682

WHO. (2023). Model list of essential medicines. https://list.essentialmeds.org/

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., Chi, E. H., & Zhou, D. (2023). Large language models as analogical reasoners, 25. https://arxiv.org/abs/2310.01714

Zhuang, Y., Liu, Q., Ning, Y., Huang, W., Lv, R., Huang, Z., Zhao, G., Zhang, Z., Mao, Q., Wang, S., & Chen, E. (2023). Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. https://arxiv.org/pdf/2306.10512.pdf

VII. Prompt used for querying the model.



Prompt template

```
"""
You are a very smart physician specializing in drugs and their efficacies and indications, when prompted with a drug fact
you will check and justify the truth value of the fact using context given to you combined with your immense knowledge
note that the stated information is merely a factual assertion.
Regarding drugs and evidence methods regarding their mechanisms of action, clinical trials and evidence based medicine.

STATED FACT:{question}. Fact check the fact with justification for your opinion using the following lines of evidence and
have a seperate section for each of these:
-Mechanisms of action (go into detail for mechanisms of action if possible)
-Evidence-based medicine
-comparisons with other treatments.
Use the given context: {context} as your foundation of knowledge for the justification.
If you cannot find the information within the given context you MUST USE YOUR VAST INTERNAL KNOWLEDGE base and go into detail don`t just state facts provide an in-depth justification
(e.g., point to mechanisms of action you are aware of, clinical trials you know) for that given section but state you used your internal knowledge base for this evidence category,
BUT DON`T HALLUCINATE if you do not know state this and do not justify using that evidence you are a medical profesional and they don't lie.

# Extra instruction: DO THE FOLLOWING FIRST!!
## Relevant Problems: Recall one example if possible of similar justification problems that are relevant to the initial problem: {question}.
# Your problem should be SIMILAR to the given context and come from your internal knowledge base (e.g., involving different treatments) and keep this to yourself but use this
# knowledge to JUSTIFY DONT EXPLAIN the stated fact dont justify your similar problem: {question}.

## Solve the Initial Problem:
Q: Copy and paste the initial problem here.
A: Explain the solution for each seperate category as listed above using the context given and if needed your internal knowledge base then
enclose the ultimate answer in \\Fact() here. In the end, one sentence to summarize your justification outcome and state the truth value for the fact here
YOU MUST GIVE A TRUTH VALUE FOR THE STATED FACT: {question} using the evidence to reason the truth of the supposed fact.
"""
```

Figure 1: Prompt

VIII. RAG workflow used within this study.

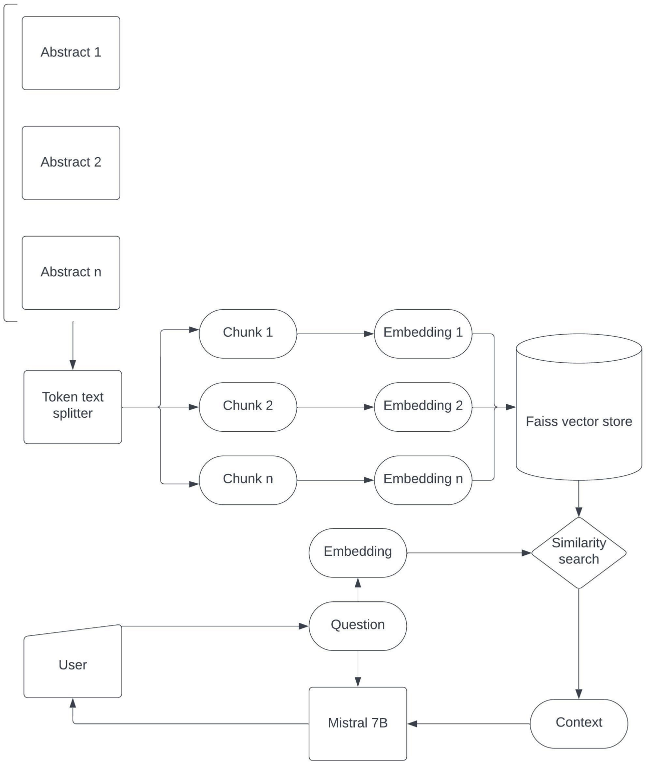
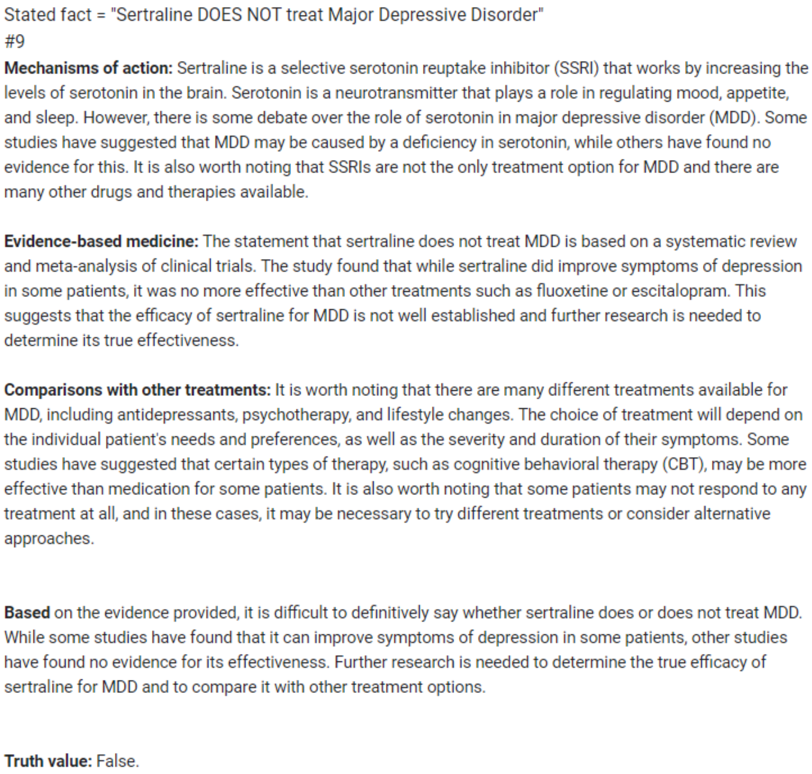


Figure 2: RAG workflow

IX. Survey sent to physician.

Here the survey sent to the physician can be observed containing the outputs generated by the LLM. https://forms.gle/AtHF1eYcAHEAEfcM6

X. Example of LLM handling the asserted fact as truth during justification and an example of correct identification in the same scenario.



Stated fact = "Sertraline DOES NOT treat Major Depressive Disorder"
#9
Mechanisms of action: Sertraline is a selective serotonin reuptake inhibitor (SSRI) that works by increasing the levels of serotonin in the brain. Serotonin is a neurotransmitter that plays a role in regulating mood, appetite, and sleep. However, there is some debate over the role of serotonin in major depressive disorder (MDD). Some studies have suggested that MDD may be caused by a deficiency in serotonin, while others have found no evidence for this. It is also worth noting that SSRIs are not the only treatment option for MDD and there are many other drugs and therapies available.

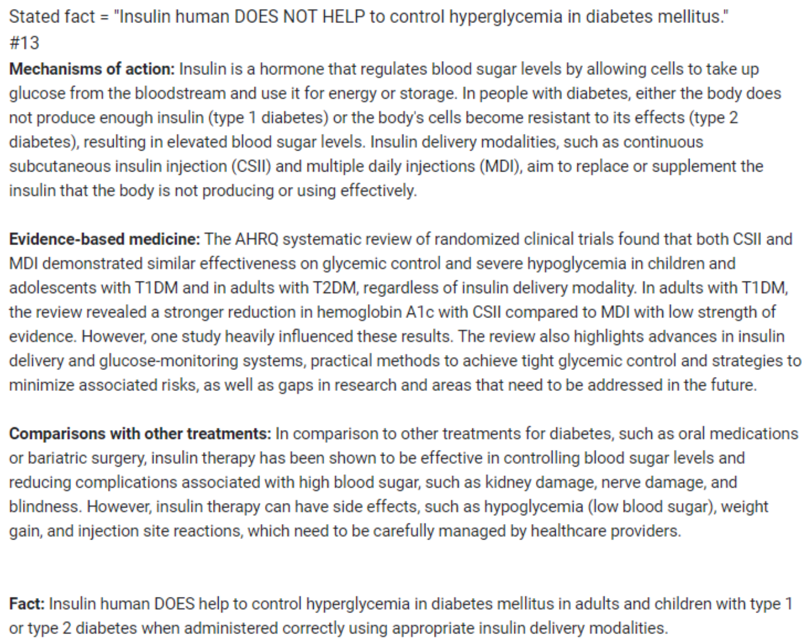
Evidence-based medicine: The statement that sertraline does not treat MDD is based on a systematic review and meta-analysis of clinical trials. The study found that while sertraline did improve symptoms of depression in some patients, it was no more effective than other treatments such as fluoxetine or escitalopram. This suggests that the efficacy of sertraline for MDD is not well established and further research is needed to determine its true effectiveness.

Comparisons with other treatments: It is worth noting that there are many different treatments available for MDD, including antidepressants, psychotherapy, and lifestyle changes. The choice of treatment will depend on the individual patient's needs and preferences, as well as the severity and duration of their symptoms. Some studies have suggested that certain types of therapy, such as cognitive behavioral therapy (CBT), may be more effective than medication for some patients. It is also worth noting that some patients may not respond to any treatment at all, and in these cases, it may be necessary to try different treatments or consider alternative approaches.

Based on the evidence provided, it is difficult to definitively say whether sertraline does or does not treat MDD. While some studies have found that it can improve symptoms of depression in some patients, other studies have found no evidence for its effectiveness. Further research is needed to determine the true efficacy of sertraline for MDD and to compare it with other treatment options.

Truth value: False.

Figure 3: Fails to identify asserted fact.



Stated fact = "Insulin human DOES NOT HELP to control hyperglycemia in diabetes mellitus."
#13

Mechanisms of action: Insulin is a hormone that regulates blood sugar levels by allowing cells to take up glucose from the bloodstream and use it for energy or storage. In people with diabetes, either the body does not produce enough insulin (type 1 diabetes) or the body's cells become resistant to its effects (type 2 diabetes), resulting in elevated blood sugar levels. Insulin delivery modalities, such as continuous subcutaneous insulin injection (CSII) and multiple daily injections (MDI), aim to replace or supplement the insulin that the body is not producing or using effectively.

Evidence-based medicine: The AHRQ systematic review of randomized clinical trials found that both CSII and MDI demonstrated similar effectiveness on glycemic control and severe hypoglycemia in children and adolescents with T1DM and in adults with T2DM, regardless of insulin delivery modality. In adults with T1DM, the review revealed a stronger reduction in hemoglobin A1c with CSII compared to MDI with low strength of evidence. However, one study heavily influenced these results. The review also highlights advances in insulin delivery and glucose-monitoring systems, practical methods to achieve tight glycemic control and strategies to minimize associated risks, as well as gaps in research and areas that need to be addressed in the future.

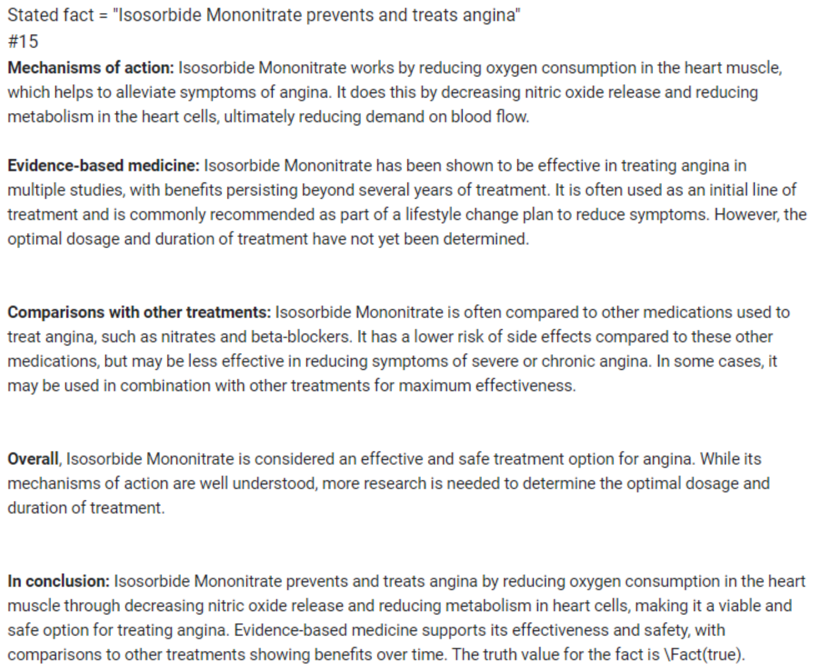
Comparisons with other treatments: In comparison to other treatments for diabetes, such as oral medications or bariatric surgery, insulin therapy has been shown to be effective in controlling blood sugar levels and reducing complications associated with high blood sugar, such as kidney damage, nerve damage, and blindness. However, insulin therapy can have side effects, such as hypoglycemia (low blood sugar), weight gain, and injection site reactions, which need to be carefully managed by healthcare providers.

Fact: Insulin human DOES help to control hyperglycemia in diabetes mellitus in adults and children with type 1 or type 2 diabetes when administered correctly using appropriate insulin delivery modalities.

Figure 4: Identifies asserted fact.

XI. Irrelevant context leads to an incorrect justification:

Here the RAG is at fault for the generation of an incorrect justification, no instance of Isosorbide was retrieved indicating its absence from the retrieved PubMed abstracts.



Stated fact = "Isosorbide Mononitrate prevents and treats angina"
#15

Mechanisms of action: Isosorbide Mononitrate works by reducing oxygen consumption in the heart muscle, which helps to alleviate symptoms of angina. It does this by decreasing nitric oxide release and reducing metabolism in the heart cells, ultimately reducing demand on blood flow.

Evidence-based medicine: Isosorbide Mononitrate has been shown to be effective in treating angina in multiple studies, with benefits persisting beyond several years of treatment. It is often used as an initial line of treatment and is commonly recommended as part of a lifestyle change plan to reduce symptoms. However, the optimal dosage and duration of treatment have not yet been determined.

Comparisons with other treatments: Isosorbide Mononitrate is often compared to other medications used to treat angina, such as nitrates and beta-blockers. It has a lower risk of side effects compared to these other medications, but may be less effective in reducing symptoms of severe or chronic angina. In some cases, it may be used in combination with other treatments for maximum effectiveness.

Overall, Isosorbide Mononitrate is considered an effective and safe treatment option for angina. While its mechanisms of action are well understood, more research is needed to determine the optimal dosage and duration of treatment.

In conclusion: Isosorbide Mononitrate prevents and treats angina by reducing oxygen consumption in the heart muscle through decreasing nitric oxide release and reducing metabolism in heart cells, making it a viable and safe option for treating angina. Evidence-based medicine supports its effectiveness and safety, with comparisons to other treatments showing benefits over time. The truth value for the fact is \Fact(true).

Figure 5: No instance of Isosorbide retrieved.