Evaluating the Medical Article Understanding Capabilities of Generative Artificial Intelligence Tools

Seyma Handan Akyon, Fatih Cagatay Akyon, Ahmet Sefa Camyar, Fatih Hizli, Talha Sari, Samil Hizli

Submitted to: JMIR Medical Informatics
on: April 07, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **24**
..... **25**
..... **26**
Figures **27**
Figure 0 ..... **28**
Multimedia Appendixes **29**
Multimedia Appendix 1 ..... **30**

Evaluating the Medical Article Understanding Capabilities of Generative Artificial Intelligence Tools

Seyma Handan Akyon[1] MD; Fatih Cagatay Akyon[2, 3] MSc; Ahmet Sefa Camyar[4] MD; Fatih Hizli[5]; Talha Sari[2, 6]; Samil Hizli[7] Prof Dr Med

[1]Golpazari Family Health Center, Bilecik, Turkey Bilecik TR
[2]SafeVideo AI San Francisco US
[3]Graduate School of Informatics, Middle East Technical University Ankara TR
[4]Cardiology Department, Ankara Etlik City Hospital Ankara TR
[5]Faculty of Medicine, Ankara, Ankara Yildirim Beyazit University Ankara TR
[6]Department of Computer Science, Istanbul Technical University Istanbul TR
[7]Department of Pediatric Gastroenterology, Ankara Bilkent City Hospital, Ankara Yildirim Beyazit University, Children Hospital Ankara TR

Corresponding Author:
Seyma Handan Akyon MD
Golpazari Family Health Center, Bilecik, Turkey
Turkey
Bilecik
TR

Abstract

Background: Reading medical articles is a challenging and time-consuming task for doctors, especially when the articles are long and complex. There is a need for a tool that can help doctors to process and understand medical articles more efficiently, accurately and fast.Generative artificial intelligence (AI) tools can assist doctors in analyzing medical articles, but there is no research evaluating medical articles and understanding the capabilities of new generative AI tools.

Objective: This study aims to critically assess and compare the comprehension capabilities of Large Language Models (LLMs) in accurately and efficiently understanding medical research articles using the STROBE checklist.

Methods: The study is a methodological type of research. The study aims to evaluate the understanding capabilities of new generative AI tools in medical articles. We designed a novel benchmark pipeline that can process PUBMED articles regardless of their length using various generative AI tools. Using this benchmark pipeline, we compared the answers of several generative AI tools (Chat-GPT 3.5-turbo, chat-GPT-4, Palm, Claude v1, Gemini pro) with the golden standard for 50 medical research articles from PUBMED. The experienced medical professor's answers to these questions are assigned as the golden standard. This study will evaluate the performance of various Large Language Models (LLMs) in accurately answering specific questions related to different sections of a scholarly article: title and abstract, methods, results, and discussion in fifteen questions from the STROBE Checklist

Results: Among the answers given by LLMs to the questions, the LLM that gave the most correct answers (66.9%) was GPT 3.5. This was followed by GPT 4-1106 version (65.6%), Palm2- (62.1%), Claude v1 (58.3%), Gemini pro (49.2%) and GPT-4 0613 version (44.1%). LLMs showcased distinct performances for each question across different parts of a scholarly article - with certain models like Palm 2 and GPT 3.5 showing remarkable versatility and depth in understanding.

Conclusions: This study is the first study to evaluate the performance of different LLMs in evaluating their ability to understanding medical articles using the Retrieval-Augmented Generation (RAG) method by giving documents.

(JMIR Preprints 07/04/2024:59258)
DOI: https://doi.org/10.2196/preprints.59258

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Research Article:

Evaluating the Medical Article Understanding Capabilities of Generative Artificial Intelligence Tools

Seyma Handan Akyon[1], Fatih Cagatay Akyon[2,3], Ahmet Sefa Çamyar[4], Fatih Hızlı[5], Talha Sarı[2,6], Şamil Hızlı[7]

[1] Gölpazarı Family Health Center, Bilecik,Turkey
[2] SafeVideo AI, San Francisco, CA, US
[3] Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
[4] Cardiology Department, Ankara Etlik City Hospital, Ankara, Turkey
[5] Faculty of Medicine, Ankara, Ankara Yıldırım Beyazıt University, Turkey
[6] Department of Computer Science, Istanbul Technical University, Istanbul, Turkey
[7] Department of Pediatric Gastroenterology, Ankara Bilkent City Hospital, Ankara Yıldırım Beyazıt University, Children Hospital, Ankara, Turkey

Corresponding author: Seyma Handan Akyon, adress: Besevler mahallesi 33.Sokak No:9 Yenimahalle/Ankara, e-mail: drseymahandan@gmail.com, phone: +905052568096

ABSTRACT

Background: Reading medical articles is a challenging and time-consuming task for doctors, especially when the articles are long and complex. A tool that can help doctors efficiently process and understand medical articles is needed.
Objectives: This study aims to critically assess and compare the comprehension capabilities of Large Language Models (LLMs) in accurately and efficiently understanding medical research articles using the STROBE checklist which provides a standardized framework for evaluating key elements of observational study.
Methods: The study is a methodological type of research. The study aims to evaluate the understanding capabilities of new generative AI tools in medical articles. A novel benchmark pipeline processed 50 medical research articles from PubMed, comparing the answers of six LLMs (GPT 3.5-turbo, GPT 4-0613, GPT 4 1106, Palm 2, Claude v1, and Gemini pro) to the benchmark established by expert medical professors. Fifteen questions, derived from the STROBE checklist, assessed LLM understanding of different sections of a research article.
Results: LLMs exhibited varying performance, with GPT 3.5-turbo achieving the highest percentage of correct answers (66.9%), followed by GPT 4-1106 (65.6%), Palm 2 (62.1%), Claude v1 (58.3%), Gemini pro (49.2%), and GPT 4-0613 (44.1%). Statistical analysis revealed statistically significant differences between LLMs ($p<0.001$), with older models showing inconsistent performance compared to newer versions. Statistical analysis revealed statistically significant differences between LLMs ($P<.001$), with older models showing inconsistent performance compared to newer versions. LLMs showcased distinct performances for each question across different parts of a scholarly article - with certain models like Palm 2 and GPT 3.5 showing remarkable versatility and depth in understanding.
Conclusions: This study is the first to evaluate the performance of different LLMs in understanding medical articles using the Retrieval-Augmented Generation (RAG) method. The findings highlight the potential of LLMs to enhance medical research by improving efficiency and facilitating evidence-based decision-making. Further research is needed to address limitations such as the influence of question formats, potential biases, and the rapid evolution of LLM models.

Keywords: *Large Language Models, ChatGPT, artificial intelligence, AI, natural language processing, generative pre-training transformer, medicine, healthcare*

INTRODUCTION

Artificial intelligence (AI) has revolutionized numerous fields, including healthcare, with its potential to enhance patient outcomes, increase efficiency, and reduce costs [1]. AI devices are divided into two main categories. Firstly, Machine Learning (ML) techniques analyze structured data

for medical applications, while the other category employs natural language processing (NLP) methods to extract information from unstructured data such as clinical notes, thereby improving the analysis of structured medical data [2]. A key development within NLP has been the emergence of large language models (LLMs), advanced systems trained on vast amounts of text data that can generate human-like language and perform a variety of language-based tasks [3, 4]. While Deep Learning models recognize patterns in data [5], LLMs are trained to predict the probability of a word sequence based on the context. By training on large amounts of text data, LLMs can generate new and plausible sequences of words that the mode has not previously observed [5]. ChatGPT (Chat Generative Pretrained Transformer), an advanced conversational AI technology developed by OpenAI in late 2022, is a general-purpose large language model [6, 7]. ChatGPT is part of a growing landscape of conversational AI products, with other notable examples including Llama (Meta), Jurassic (Ai21), Claude (Anthropic), Command (Cohere), Gemini, Palm, and Bard (Google) [6,7]. The potential of AI systems to enhance medical care and health outcomes is highly promising [8]. Therefore, it is essential to ensure that the creation of AI systems in healthcare adheres to the principles of trust and explainability. Evaluating the medical knowledge of AI systems compared to that of expert clinicians is a vital initial step to assess these qualities [9–11].

Reading medical articles is a challenging and time-consuming task for doctors, especially when the articles are long and complex. This poses a significant barrier to efficient knowledge acquisition and evidence-based decision making in healthcare. There is a need for a tool that can help doctors to process and understand medical articles more efficiently and accurately. Although LLMs are promising in evaluating patients, diagnosis, and treatment processes [13], studies on reading academic articles are limited. LLMs can be directly questioned and can generate answers from their own memory [14,15]. This has been extensively studied in many articles. However, these pose the problem of artificial hallucinations, which are inaccurate outputs, in LLMs. The retrieval augmented generation (RAG) method, which intuitively addresses the knowledge gap by conditioning language models on relevant documents retrieved from an external knowledge source, can be used to overcome this issue [16].

The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist provides a standardized framework for evaluating key elements of observational study and sufficient information for critical evaluation. These guidelines consist of 22 items that authors should adhere to before submitting their manuscripts for publication [17, 18, 19]. This study aims to address this gap by evaluating the comprehension capabilities of LLMs in accurately and efficiently understanding medical research articles. We utilize the The STROBE checklist to assess LLMs' ability to understand different sections of research articles. Our study employs a novel benchmark pipeline that can process PubMed articles regardless of their length using various generative AI tools. This research will provide critical insights into the strengths and weaknesses of different LLMs in enhancing medical research article comprehension. To overcome the problem of "artificial hallucinations," we implement the RAG method. RAG involves providing the LLMs with a prompt that instructs them to answer while staying relevant to the given document, ensuring responses align with the provided information. The results of our study will provide valuable information for medical professionals, researchers, and developers seeking to leverage the potential of LLMs for improving medical literature comprehension and ultimately enhance patient care and research efficiency.

METHODS

Design of Study

This study employs a methodological research design to evaluate the comprehension capabilities of

generative AI tools using the STROBE checklist.

Article Selection

We included the first 50 observational studies conducted within the past five years that were retrieved through an advanced search on PubMed on December 19, 2023, using "obesity" in the title as the search term. The included studies were limited to those written in English, available as free full-text, and focusing specifically on human subjects **(Figure 1).** The articles included in the study were statistically examined in detail, and a total of 11 of them were excluded because they were not observational studies. The study was completed with 39 articles. A post-hoc power analysis was conducted to assess the statistical power of our study based on the total correct responses across all repetitions. The analysis excluded ChatGPT 4-1106 and ChatGPT 3.5 Turbo-1106 (31.8%) due to their similar performance and the significant differences observed between other models. The power analysis, conducted using GPower, indicated that all analyses exceeded 95% power. Thus, the study was completed with the 39 selected articles, ensuring sufficient statistical power to detect meaningful differences in LLM performance.

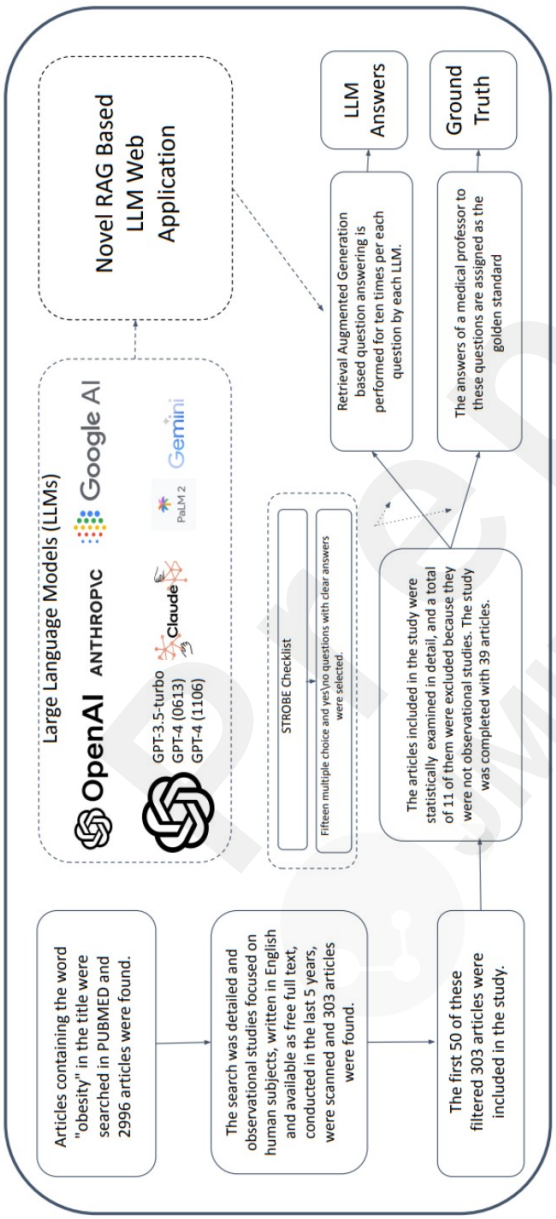


Figure 1. Flow Chart: Recruitment and Data Collection Process for Evaluating LLM Comprehension of Medical Research Articles

Benchmark Development

Our study employed a novel benchmark pipeline to evaluate the understanding capabilities of LLMs when processing medical research articles. To establish a reference standard for evaluating the LLMs' comprehension, we relied on the expertise of an experienced medical professor and an epidemiology expert doctor. The professor, with their extensive medical knowledge, was tasked with answering 15 questions derived from the STROBE checklist, designed to assess key elements of observational studies and covering different sections of a research article (Table 1). The epidemiology expert doctor, with their specialized knowledge in statistical analysis and epidemiological methods, provided verification and validation of the professor's answers, ensuring the rigor of the benchmark. The combined expertise of both professionals provided a robust and reliable reference standard against which the LLMs' responses were compared.

Table 1. The questions derived from the STROBE checklist for observational study and answers

Questions	Answers
Title and abstract	
Q1. Does the article indicate the study's design with a commonly used term in the title or the abstract?	1: yes 2: no
Methods	
Q2. What is the observational study type: cohort, case-control, or cross-sectional studies?	1: cohort study 2: a case-control study 3: cross-sectional study 4: The study type is not stated in the article
Q3. Were settings/ locations mentioned in the method?	1: yes 2: no
Q4. Were relevant dates mentioned in the method?	1: yes 2: no
Q5. Were eligibility criteria for selecting participants mentioned in the method?	1: yes 2: no
Q6. Were sources and methods of selection of participants mentioned in the method?	1: yes 2: no
Q7. Were any efforts to address potential sources of bias described in the method or discussion?	1: yes 2: no
Q8. Which program was used for statistical analysis?	1: SPSS (Statistical Package for Social Sciences) was used for statistical analysis. 2: MedCalc was used for statistical analysis. 3: SAS (Statistical Analysis System) was used for statistical analysis. 4: STATA was used for statistical analysis. 5: R program was used. 6: Another program was used for statistical analysis. 7: The program for statistical analysis is not specified.
Results	
Q9. Were report numbers of individuals at each stage of the study (e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed) mentioned in the results?	1: yes 2: no
Q10. Was a flowchart used to show the reported numbers of individuals at each stage of the study?	1: yes 2: no
Q11. Were the study participants' demographic characteristics (e.g. age, sex) given in the results?	1: yes 2: no
Discussion	
Q12. Does the discussion part summarize key results concerning study objectives?	1: yes 2: no
Q13. Are the limitations of the study discussed in the article?	1: yes

	2: no
Q14. Is the generalisability of the study discussed in the discussion part?	1: yes 2: no
Funding	
Q15. Is the funding of the study mentioned in the article?	1: yes 2: no

This list of fifteen questions, two multiple-choice, and thirteen yes/no questions has been prepared by selecting the STROBE Checklist items that can be answered definitively and have clear, non-subjective responses. Q1, related to title and abstract, examines the LLMs' ability to identify and understand research designs and terms that are commonly used, evaluating the model's comprehension of the concise language typically used in titles and abstracts. Q2-Q8, related to methods, covers various aspects of the study's methodology, from the type of observational study to the statistical analysis programs used. They test the model's understanding of the detailed and technical language often found in this section. Q9-Q11, related to results, focuses on the accuracy and completeness of reported results, such as participant numbers at each study stage and demographic characteristics. These questions gauge the LLMs' capability to parse and summarize factual data. Q12-Q14, related to the discussion, involves summarizing key results, discussing limitations, and the study's generalizability. These questions assess the LLMs' ability to engage with more interpretive and evaluative content, showcasing its understanding of research impacts and contexts. Q15, related to funding, tests the LLMs' attentiveness to specific yet crucial details that could influence the interpretation of research findings.

Development of Novel RAG Based LLM Web Application

The methodology incorporated a novel web application specifically designed for this purpose to assess the understanding capabilities of generative AI tools in medical research articles (Figure 2). To mitigate the problem of "artificial hallucinations" inherent to LLMs, our study implemented the RAG method, which involves using a web application to dissect PDF-format medical articles from PubMed into text chunks ready to be processed by various LLMs. This approach guides the LLMs to provide answers grounded in the provided information by supplying them with relevant text chunks retrieved from the target article.

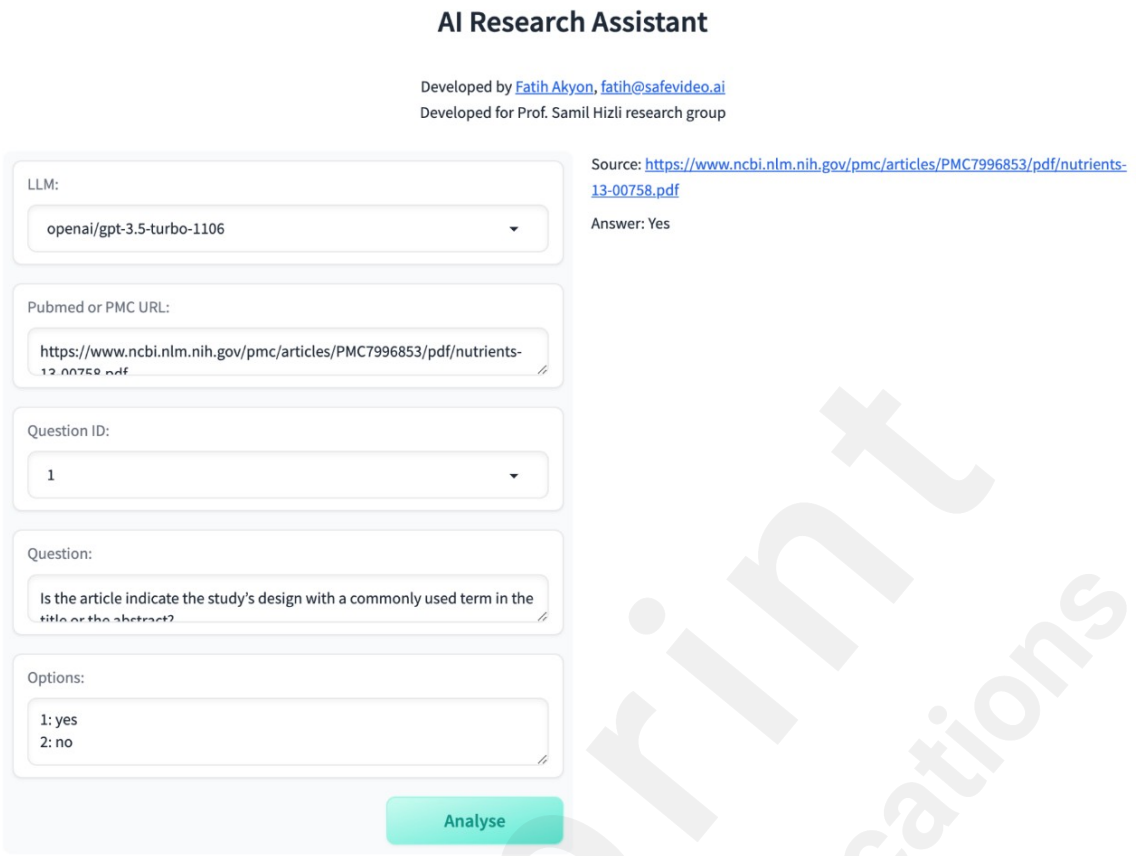


Figure 2. Novel RAG Based LLM Web Application Interface

Benchmark Pipeline

The benchmark pipeline itself is designed to process PubMed articles of varying lengths and extract relevant information for analysis. This pipeline operates as follows:

- Article Retrieval: We retrieved 39 observational studies from PubMed using the search term "obesity" in the title.
- Text Extraction and Chunking: Each retrieved PubMed article was converted to PDF format and then processed through our web application. The application extracts all text content from the article and divides it into smaller text chunks of manageable size.
- Vector Representation: Using the OpenAI text-ada-embedding-002 model, each text chunk was converted into a representation vector. These vectors capture the semantic meaning of the text chunks, allowing for efficient information retrieval.
- Vector Database Storage: The generated representation vectors were stored in a vector database (LanceDB in our case). This database allows for rapid searching and retrieval of the most relevant text chunks based on a given query.
- Query Processing: When a query (question from the STROBE checklist) was posed to an LLM, our pipeline calculated the cosine similarities between the query's representation vector and the vectors stored in the database. This identified the most relevant text chunks from the article.
- Retrieval-Augmented Generation: The retrieved text chunks, along with the original query, were then combined and presented to the LLM. This approach, known as RAG, ensured that the LLM's responses were grounded in the specific information present in the article, mitigating the risk of hallucinations.

- Answer Generation and Evaluation: The LLM generated an answer to the query based on the provided text chunks. The accuracy of each LLMs' response was then evaluated by comparing it to the benchmark answers provided by a medical professor.

LLMs

Using this benchmark pipeline, we compared the answers of the generative AI tools, which are ChatGPT 3.5-turbo 1106 (11th June version), ChatGPT 4-0613 (6th November version), ChatGPT 4-1106 (11th June version), Palm 2 (chat-bison), Claude v1, Gemini pro with the benchmark in 15 questions for 39 medical research articles **(Table 2)**. In this study, 15 questions selected from the STROBE checklists were posed 10 times each for 39 articles to six different LLMs.

Access issues with Claude v1, specifically restrictions on its ability to process certain medical information, resulted in the exclusion of data from six articles, limiting the study's scope to 33 articles. LLMs commonly provide a "knowledge-cutoff" date, indicating the point at which their training data ends and they may not have access to the most up-to-date information. With some LLMs, however, the company doesn't explicitly state a cutoff date. That's why the explicitly stated cutoff dates are given in Table 2, based on the publicly available information for each LLM.

Table 2 . The generative AI tools compared with the benchmark in study

Generative-AI Tool	Version	Company	Cutoff Date
GPT 3,5-turbo	6th November 2023	OpenAI	September 2021
GPT 4-0613	13th June 2023	OpenAI	September 2021
GPT 4-1106	6th November 2023	OpenAI	April 2023
Claude v1	version 1	Anthropic	*
Palm 2	chat-bison	Google	*
Gemini pro	1.0	Google	*

**The company doesn't explicitly state a cutoff date.*

A chatbot session begins when a user inputs a system query, referred to as a system prompt in everyday language, with the chatbot responding in natural language within about a second, facilitating an interactive conversation-like exchange enabled by the system's contextual awareness. In addition to the RAG method, by providing LLMs with convenient system prompts that instruct them to answer while staying relevant to the given document, it is possible to generate responses that align with the provided information. We used the following system prompt for all LLMs:

"You are an expert medical professor specialized in pediatric gastroenterology hepatology and nutrition, with a detailed understanding of various research methodologies, study types, ethical considerations, and statistical analysis procedures. Your task is to categorize research articles based on information provided in query prompts. There are multiple options for each question, and you must select the most appropriate one based on your expertise and the context of the research article presented in the query."

The language models used in this study rely on statistical models that incorporate random seeds to facilitate the generation of diverse outputs. However, the companies behind these LLMs do not offer a stable way to fix these seeds, meaning that a degree of randomness is inherent in their responses. To further control this randomness, we utilized the "temperature" parameter within the language models. This parameter allows for adjustment of the level of randomness, with a lower temperature setting generally producing more deterministic outputs. For this study, we opted for a low-temperature parameter setting of 0.1 to minimize the impact of randomness. Despite these efforts, complete elimination of randomness is not possible. To further mitigate its effects and enhance the consistency of our findings, we repeated each question ten times for the same language model. By analyzing the responses across these ten repetitions, we could determine the frequency of accurate and consistent answers. This approach helped to identify instances where the LLM's responses were consistently aligned with the benchmark answers, highlighting areas of strength and consistency in comprehension.

Statistical Analysis

Each question was repeated ten times repetitively in the same time period to obtain answers from multiple LLMs and ensure the consistency and reliability of responses. Consequently, the responses to the same question were analyzed to determine how many aligned with the benchmark, and the findings were examined. Only the answers that were correct and followed the instructions provided in the question text were considered "correct". Ambiguous answers, evident mistakes, and responses with an excessive number of candidates were considered incorrect. The data was carefully examined, and the findings were documented and analyzed. Each inquiry and its response formed the basis of the analysis. Various descriptive statistical tests were used to assess the data presented as numbers and percentages. The Shapiro-Wilk test was used to assess the data's normal distribution. The Kruskal-Wallis and Pearson chi-square tests were employed in the statistical analysis. Type I error level was accepted as 5% in the analyses performed using the SPSS 29.0.

Ethical Considerations

This study only used information that had already been published on the Internet. Ethical approval was not required for this study since it did not involve any human or animal research participants. This study did not involve a clinical trial as it focused on evaluating the capabilities of AI tools in understanding medical articles.

RESULTS

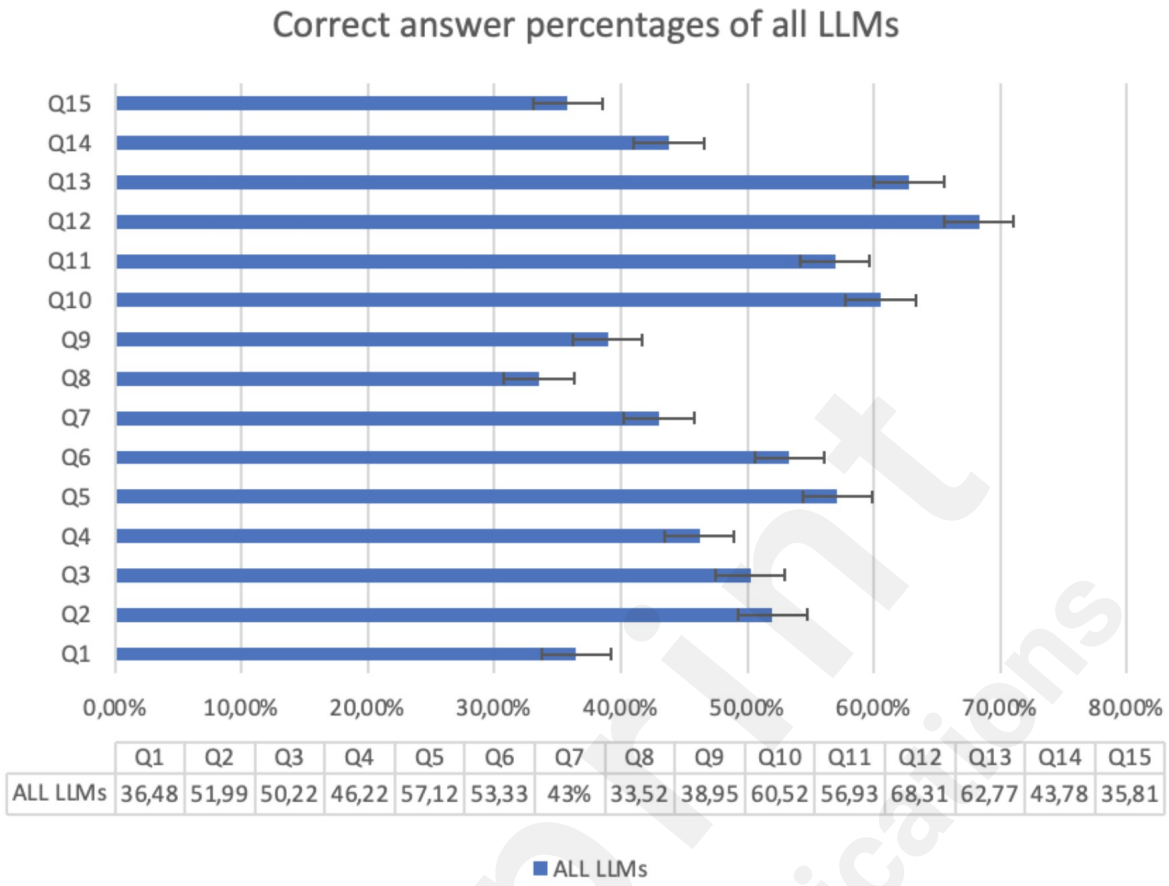
In this study, 15 questions selected from the STROBE checklists were posed 10 times each for 39 articles to six different LLMs. Access issues with Claude v1, specifically restrictions on its ability to process certain medical information, resulted in the exclusion of data from six articles, limiting the study's scope to 33 articles. The percentage of correct answers for each LLM is shown in **Table 3**, with GPT 3.5-turbo achieving the highest rate (66.9%), followed by GPT 4-1106 (65.6%), Palm 2 (62.1%), Claude v1 (58.3%), Gemini pro (49.2%), and GPT 4-0613 (44.1%).

Table 3. Comparison of Correct Answers Among LLMs

LLM	Total questions asked	True answers (n)	True answers (%)
ChatGPT 3.5 Turbo-1106	5850	3916	**66.9**
ChatGPT 4-0613	5850	2580	44.1
ChatGPT 4-1106	5850	3837	**65.6**
Claude v1	4950	2887	58.3
Palm2-chatbison	5850	3632	**62.1**
Gemini pro	5850	2878	49.2

Each LLM was compared with another LLM that provided a lower percentage of correct answers. Statistical analysis using the Kruskal-Wallis test revealed statistically significant differences between the LLMs (P<.001). The lowest correct answer percentage was provided by ChatGPT 4–0613, at 44.1%. Gemini pro yielded 49.2% correct answers, significantly higher than Chat-GPT 4 – 0613 (P<.001). Claude v1 yielded 58.3% correct answers, statistically significantly higher than Gemini Pro (P<.001). Palm 2 achieved 62.1% correct answers, significantly higher than Claude v1 (P<.001). ChatGPT 4-1106 achieved 65.6% correct answers, significantly higher than Palm 2 (P<.001). The difference between ChatGPT 4-1106 and ChatGPT 3.5 Turbo-1106 was not statistically significant (P=.061). Of the 39 articles analyzed, 28 (71.8%) were published before the training data cutoff date for GPT-3.5-turbo and GPT-4-0613, while all 39 articles (100%) were published before the cutoff date for GPT-4-1106. Explicit cutoff dates for the remaining LLMs (Claude, Palm 2, and Gemini Pro) were not publicly available and therefore could not be assessed in this study.

When all LLMs are collectively considered, the three questions receiving the highest percentage of correct answers were Q12 (68.31%), Q13 (62.77%), and Q10 (60.52%). Conversely, the three questions with the lowest percentage of correct responses were Q8 (33.52%), Q15 (35.81%), and Q1 (36.48%). **(Figure 3)**



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
ALL LLMs	36,48	51,99	50,22	46,22	57,12	53,33	43%	33,52	38,95	60,52	56,93	68,31	62,77	43,78	35,81

Figure 3. Correct answer percentages of LLMs

The percentages of correct answers given by all LLMs for each question are depicted in **Table 4**. The median values for questions 7, 8, 9, 10, and 14 were similar across all LLMs, indicating a general consistency in performance for these specific areas of comprehension. However, significant differences were observed in the performance of different LLMs for other questions. The statistical tests used in this analysis were the Kruskal-Wallis test for comparing the medians of multiple groups and the chi-square test for comparing categorical data. For question 1, the fewest correct answers were provided by Claude (24.8%) and Gemini pro (39.5%), while the most correct answers were provided by Palm 2 (60.3%) (P=.011). In question 2, Claude v1 (73.3%) achieved the highest median correct answer count (10.0), while Gemini Pro provided the fewest correct answers (47.4%) (P=.028). For question 3, GPT 3.5 (85.1%) and Palm 2 (86.8%) had the highest median correct answer counts, while GPT 4-0613 (32.8%) and Gemini Pro (37.9%) had the lowest (P<.001). In the fourth question, Palm 2 (73.8%), GPT 3.5 (58.7%), and GPT 4-1106 (67.2%) performed best, while GPT 4-0613 (37.4%) showed the lowest performance (P<.001). For questions 5 and 6, GPT 4-0613 (41.8%) and Gemini Pro (37.2%) provided fewer correct answers compared to the other LLMs (P<.001 and P=.001, respectively). In question 11, GPT 4-1106 (81.3%), Claude (69.4%), and Palm 2 (81.2%) performed well, while Gemini Pro (52.8%) had the fewest correct answers (P=.001). For questions 12 and 13, ll LLMs, except GPT 4-0613, performed well in these areas (P<.001). In question 15, GPT 3.5 (73.6%) showed the highest number of correct answers (P<.001). (Multimedia Appendix 2)

Table 4. Comparative Analysis of Correct Responses by LLMs Across Ten Iterations for Each Question

Questions	1	2	3	4	5	6	P
Title and abstract (Q1)							
Q1	48.5% 4.0 (3.0-8.0)	46.9% 5.0 (0.0-10.0)	37.7% 2.0 (0.0-7.0)	24.8% 1.0 (0.0-4.0)	60.3% 7.5 (1.8-10.0)	39.5% 2.0 (0.0-9.0)	.011
Methods (Q2-8)							
Q2	62.8% 8.0 (2.0-10.0)	53.8% 6.0 (0.0-10.0)	70.0% 9.0 (4.0-10.0)	73.3% 10.0 (5.0-10.0)	68.5% 8.0 (4.8-10.0)	47.4% 4.0 (0.0-10.0)	.028
Q3	85.1% 10.0 (8.0-10.0)	32.8% 2.0 (0.0-6.0)	71.0% 8.0 (5.0-10.0)	48.8% 5.0 (2.0-7.5)	86.8% 10.0 (9.0-10.0)	37.9% 3.0 (0.0-7.0)	<.001
Q4	58.7% 7.0 (3.0-10.0)	37.4% 3.0 (0.0-7.0)	67.2% 7.0 (4.0-10.0)	50.3% 5.0 (0.0-10.0)	73.8% 9.0 (5.8-10.0)	46.2% 4.0 (0.0-8.0)	<.001
Q5	84.7% 10.0 (9.0-10.0)	41.8% 4.0 (0.0-8.0)	86.4% 10.0 (8.0-10.0)	80.9% 10.0 (8.0-10.0)	83.2% 10.0 (9.5-10.0)	37.2% 3.0 (0.0-7.0)	<.001
Q6	70.0% 10.0 (3.0-10.0)	49.5% 7.0 (0.0-9.0)	74.1% 10.0 (5.0-10.0)	72.1% 10.0 (2.5-10.0)	69.1% 10.0 (0.0-10.0)	50.3% 6.0 (0.0-9.0)	.001
Q7	47.7% 4.0 (0.0-10.0)	46.7% 4.0 (1.0-9.0)	46.7% 5.0 (0.0-10.0)	59.7% 8.0 (0.0-10.0)	60.9% 10.0 (0.0-10.0)	49.7% 5.0 (0.0-10.0)	.553
Q8	46.7% 4.0 (0.0-9.0)	32.3% 3.0 (0.0-5.0)	41.0% 4.0 (1.0-7.0)	31.8% 3.0 (0.0-5.5)	50.6% 5.0 (0.0-10.0)	38.5% 3.0 (0.0-8.0)	.351
Results (Q9-11)							
Q9	41.0% 0.0 (0.0-10.0)	46.7% 4.0 (0.0-10.0)	50.5% 5.0 (3.0-7.0)	32.4% 0.0 (0.0-10.0)	63.8% 8.5 (2.8-10.0)	44.1% 0.0 (0.0-10.0)	.053
Q10	72.3% 9.0 (4.0-10.0)	74.1% 10.0 (5.0-10.0)	78.2% 10.0 (4.0-10.0)	72.1% 10.0 (2.0-10.0)	66.2% 9.0 (3.0-10.0)	71.0% 10.0 (4.0-10.0)	.625
Q11	72.8% 8.0 (5.0-10.0)	53.3% 6.0 (1.0-10.0)	81.3% 10.0 (7.0-10.0)	69.4% 10.0 (1.0-10.0)	81.2% 10.0 (8.5-10.0)	52.8% 4.0 (2.0-10.0)	<.001
Discussion (Q12-14)							
Q12	80.0% 10.0 (6.0-10.0)	56.4% 6.0 (3.0-10.0)	93.6% 10.0 (10.0-10.0)	89.1% 10.0 (10.0-10.0)	88.8% 10.0 (10.0-10.0)	84.9% 10.0 (8.0-10.0)	<.001
Q13	98.5% 10.0 (10.0-10.0)	22.8% 1.0 (0.0-4.0)	96.9% 10.0 (10.0-10.0)	81.2% 10.0 (7.0-10.0)	86.5% 10.0 (9.8-10.0)	67.4% 10.0 (3.0-10.0)	<.001
Q14	61.8% 8.0 (2.0-10.0)	39.5% 0.0 (0.0-10.0)	49.5% 4.0 (2.0-10.0)	55.8% 6.0 (2.5-10.0)	62.1% 9.0 (1.5-10.0)	47.7% 4.0 (1.0-8.0)	.151
Funding (Q15)							
Q15	73.6% 9.0 (4.0-10.0)	27.4% 0.0 (0.0-4.0)	39.7% 3.0 (0.0-8.0)	33.0% 3.0 (0.0-6.0)	60.9% 7.0 (3.0-10.0)	23.3% 0.0 (0.0-3.0)	<.001

1=GPT 3.5 Turbo-1106, 2=GPT 4-0613 3=GPT 4-1106 4=Claude v1 5=Palm2-chat bison, 6=Gemini pro
Each cell in the table displays the percentage of accurate answers, the minimum and maximum correct answer counts, and the median correct answer count out of ten trials. The analysis was conducted using the Kruskal-Wallis test.

0-25.0%	25.1-50.0%	50.1-75.0%	75.1-100.0%

DISCUSSION

AI can improve the data analysis and publication process in scientific research while also being used to generate medical articles [20]. Although these fraudulent articles may appear well-crafted, their semantic inaccuracies and errors can be detected by expert readers upon closer examination [15,21]. The impact of LLMs on healthcare is often discussed in terms of their ability to replace health professionals, but their significant impact on medical and research writing applications and limitations is often overlooked. Therefore, physicians involved in research need to be cautious and verify information when using LLMs. As their reliance can lead to ethical concerns and inaccuracies, the scientific community should be vigilant in ensuring the accuracy and reliability of AI tools by using them as aids rather than replacements, understanding their limitations and biases [22,23]. With millions of papers published annually, AI could generate summaries or recommendations, simplifying the process of gathering evidence and enabling researchers to grasp important aspects of scientific results more efficiently [23]. Moreover, there is limited research focused on assessing the comprehension of academic articles.

This study aimed to evaluate the ability of six different LLMs to understand medical research articles using the STROBE checklist. We employed a novel benchmark pipeline that processed 39 PubMed articles, posing 15 questions derived from the STROBE checklist to each model. The benchmark was established using the answers provided by an experienced medical professor and validated by an epidemiologist, serving as a reference standard against which the LLMs' responses were compared. To mitigate the problem of "artificial hallucinations" inherent to LLMs, our study implemented the RAG method, which involves using a web application to dissect PDF-format medical articles into text chunks and present them to the LLMs.

Our findings reveal significant variation in the performance of different LLMs, suggesting that LLMs are capable of understanding medical articles to varying degrees. While newer models like GPT 3.5 Turbo and GPT 4-1106 generally demonstrated better comprehension, GPT 3.5 Turbo outperformed even the more recent GPT 4-0613 in certain areas. This unexpected finding highlights the complexity of LLM performance, indicating that simple assumptions about newer models consistently outperforming older ones may not always hold true. The impact of training data cutoffs on LLM performance is a critical consideration in evaluating their ability to understand medical research [24]. While we were able to obtain explicitly stated cutoff dates for GPT-3.5-turbo, GPT-4-1106 and GPT-4-0613, this information was not readily available for the remaining models. This lack of transparency regarding training data limits our ability to definitively assess the impact of knowledge cutoffs on model performance. The observation that all 39 articles were published before the cutoff date for GPT-4-1106, while only 28 articles were published before the cutoff date for GPT-3.5-turbo and GPT-4-0613, suggests that the knowledge cutoff may play a role in the observed performance differences. GPT-4-1106, with a more recent knowledge cutoff, has access to a larger dataset, potentially including information from more recently published research. This could contribute to its generally better performance compared to GPT-3.5-turbo. However, it's important to note that GPT-3.5-turbo still outperformed GPT-4-0613 in specific areas, even with a similar knowledge cutoff. This suggests that factors beyond training data, such as (e.g., the number of layers, the type of attention mechanism, or the use of transformers) and compression techniques (e.g., quantization, pruning, or knowledge distillation), may also play a significant role in LLM performance. Future research should prioritize transparency regarding training data cutoffs and aim to standardize how LLMs communicate these crucial details to users.

This study evaluated the performance of various LLMs in accurately answering specific questions related to different sections of a scholarly article: title and abstract, methods, results,

discussion, and funding. The results shed light on which LLMs excel in specific areas of comprehension and information retrieval from academic texts. Palm 2 (60.3%) showed superior performance in Q1, identifying the study design from the title or abstract, suggesting enhanced capability in understanding and identifying specific terminologies. Claude (24.8%) and Gemini pro (39.5%), however, lagged, indicating a potential area for improvement in terminology recognition and interpretation. Claude v1 (73.3%) and Palm 2 (86.8%) exhibited strong capabilities in identifying methodological details, such as observational study types and settings/locations (Q2-Q3). This suggests a robust understanding of complex methodological descriptions and the ability to distinguish between different study frameworks. For questions regarding the results section (Q9-Q11), it's evident that models like GPT 4-1106 (81.3%), Claude (69.4%), and Palm 2 (81.2%) showed superior performance in providing correct answers related to the study participants' demographic characteristics and the use of flow charts. All LLMs except for GPT4-0613 (22.8%) exhibited remarkable competence in summarizing key results, discussing limitations, and addressing the generalizability of the study (Q12-Q14), which are critical aspects of the discussion section. GPT 3.5 (73.6%) particularly excelled in identifying the mention of funding (Q15), indicating a nuanced understanding of acknowledgments and funding disclosures often nuanced and embedded towards the end of articles. Across the array of tested questions, both GPT 3.5 and Palm 2 exhibit remarkable strengths in understanding and analyzing scholarly articles, with Palm 2 generally showing a slight edge in versatility, especially in interpreting methodological details and study design. GPT 3.5, while strong in discussing study limitations, generalized findings, and funding details, indicates that improvements can be made in extracting complex methodological information. We observed that different models excelled in different areas, indicating that no single LLM currently demonstrates universal dominance in medical article understanding. This suggests that factors like training data, model architecture, and question complexity influence performance, and further research is needed to understand the specific contributions of each factor.

LLMs can be directly questioned and can generate answers from their own memory [15]. This has been extensively studied in many medical articles. According to a study, ChatGPT, a large language model, was evaluated on the United States Medical Licensing Exam (USMLE). The results showed that ChatGPT performed at or near the passing threshold for exams without any specialized training, demonstrating a high level of concordance and insight in its explanations. These findings suggest that LLMs have the potential to aid in medical education and potentially assist with clinical decision-making [10, 25]. Another study aimed to evaluate the knowledge level of ChatGPT in medical education by assessing its performance in a multiple-choice question examination and its potential impact on the medical examination system. The results indicated that ChatGPT achieved a satisfactory score in both basic and clinical medical sciences, highlighting its potential as an educational tool for medical students and faculty [26]. Furthermore, ChatGPT offers information and aids healthcare professionals in diagnosing patients by analyzing symptoms and suggesting appropriate tests or treatments. However, advancements are required to ensure AI's interpretability and practical implementation in clinical settings [11]. The study conducted in October 2023 explored the diagnostic capabilities of GPT-4V, an AI model, in complex clinical scenarios involving medical imaging and textual patient data. Results showed that GPT-4V had the highest diagnostic accuracy when provided with multimodal inputs, aligning with confirmed diagnoses in 80.6% of cases [27]. In another study, GPT-4 was instructed to address the case with multiple-choice questions followed by an unedited clinical case report that evaluated the effectiveness of the newly developed AI model GPT-4 in solving complex medical case challenges. Correctly diagnosed 57% of cases, beating 99.98% of computer-generated human readers [28]. These studies highlight the potential of multimodal AI models like GPT-4 in clinical diagnostics, but further investigation is needed to uncover biases and limitations due to the model's proprietary training data and architecture.

There are few studies in which LLMs are directly questioned, and their capacities to produce answers from their own memories are compared with each other and expert clinicians. In a study, ChatGPT-3.5 and GPT-4 were compared to orthopedic residents in their performance on the American Board of Orthopaedic Surgery written examination., with residents scoring higher overall, and a subgroup analysis revealed that ChatGPT-3.5 and ChatGPT-4 outperformed residents in answering text-only questions, while residents scored higher in image interpretation questions. GPT-4 scored higher than GPT-3.5 [29]. A study aimed to evaluate and compare the recommendations provided by GPT-3 and ChatGPT-4 with those of primary care physicians for the management of depressive episodes. The results showed that both GPT-3.5 and GPT-4 largely aligned with accepted guidelines for treating mild and severe depression while demonstrating a lack of gender or socioeconomic biases observed among primary care physicians. However, further research is needed to refine the AI recommendations for severe cases and address potential ethical concerns and risks associated with their use in clinical decision-making [30]. Another study assessed the accuracy and comprehensiveness of health information regarding urinary incontinence generated by various LLMs. By inputting selected questions into GPT-3.5, GPT-4, and BARD, the researchers found that GPT-4 performed the best in terms of accuracy and comprehensiveness, surpassing GPT-3.5 and BARD [31]. According to a study evaluates the performance of two ChatGPT models (GPT-3.5 and GPT-4) and human professionals in answering ophthalmology questions from the StatPearls question bank, GPT-4 outperformed both GPT-3.5 and human professionals on most ophthalmology questions, showing significant performance improvements and emphasizing the potential of advanced AI technology in the field of ophthalmology [32]. Some studies showed that GPT-4 is more proficient, as evidenced by scoring higher than ChatGPT 3.5 in both multiple-choice dermatology exams and non-multiple-choice cardiology heart failure questions from various sources and outperforming GPT-3.5 and Flan-PaLM 540B on medical competency assessments and benchmark datasets [33-35]. In a study conducted on the proficiency of various open-source and proprietary large language models in the context of nephrology multiple-choice test-taking ability, it was found that their performance on 858 nephSAP questions ranged from 17.1% to 30.6%, with Claude 2 at 54.4% accuracy and GPT-4 at 73.3%, highlighting the potential for adaptation in medical training and patient care scenarios [36]. To our knowledge, this is the first study to assess the performance of evaluating medical articles and understanding the capabilities of different LLMS. The findings reveal that the performance of LLMs varies across different questions, with some LLMs showing superior understanding and answer accuracy in certain areas. Comparative analysis across different LLMs showcases a gradient of capabilities. The results revealed a hierarchical performance ranking as follows: GPT 4 -1106 equals GPT 3.5 Turbo, which is superior to Palm 2, followed by Claude v1, then Gemini Pro, and lastly, GPT 4 0613. Similar to the literature review, GPT-4 1106 and GPT 3.5 showed improved accuracy and understanding compared to other LLMs. This mirrors wider literature trends, indicating LLMs' rapid evolution and increasing sophistication in handling complex medical queries. Notably, GPT 3.5 Turbo showed better performance than GPT 4 0613, which may be counterintuitive considering the tendency to assume newer iterations naturally perform better. This anomaly in performance between newer and older versions can be attributed to the application of compression techniques in developing new models to reduce computational costs. While these advancements make deploying LLMs more cost-effective and thus accessible, they can inadvertently compromise the performance of LLMs. The notable absence of responses from Palm in certain instances, actually stemming from Google's policy to restrict the usage of its medical information, presents an intriguing case within the scope of our discussion. Despite these constraints, Palm's demonstrated high performance in other areas is both surprising and promising. This suggests that even when faced with limitations on accessing a vast repository of medical knowledge, Palm's underlying architecture and algorithms enable it to make effective use of the information it can access, showcasing the robust potential of LLMs in medical settings even under restricted conditions.

While LLMs can be directly questioned and generate answers from their own memory, as demonstrated in numerous studies above, this approach can lead to inaccuracies known as hallucinations. Hallucinations in LLMs have diverse origins, encompassing the entire spectrum of the capability acquisition process, with hallucinations primarily categorized into three aspects: training, inference, and data. Architecture flaws, exposure bias, and misalignment issues in both pre-training and alignment phases induce hallucinations. To address this challenge, our study utilized the RAG method, ensuring that the LLMs' responses were grounded in factual information retrieved from the target article. The RAG method intuitively addresses the knowledge gap by conditioning language models on relevant documents retrieved from an external knowledge source [16,37]. RAG provides the LLM with relevant text chunks extracted from the specific article being analyzed. This ensures that the LLM's responses are directly supported by the provided information, reducing the risk of hallucination. While a few studies have explored the use of RAG to compare LLMs, like the one demonstrating GPT-4's improved accuracy with RAG for interpreting oncology guidelines [38], our study is the first to evaluate LLM comprehension of medical research articles using this method. This method conditions LLMs on relevant documents retrieved from an external knowledge source, ensuring their answers are grounded in factual information. Moreover, the system prompt's design is crucial for LLMs' output, incorporating background context, task instructions, and formatting constraints [39]. In this study, it is empirically determined that a foundational system and set of system prompts universally enhanced the response quality across all language models tested. This approach was designed to optimize the comprehension and summarization capabilities of each generative AI tool when processing medical research articles. The specific configuration of system settings and query structures we identified significantly contributed to improving the accuracy and relevance of the models' answers. These optimized parameters were crucial in achieving a more standardized and reliable evaluation of each model's ability to understand complex medical texts. While further research is needed to fully understand the effectiveness of RAG across different medical scenarios, our findings demonstrate its potential to enhance the reliability and accuracy of LLMs in medical research comprehension.

This study, while offering valuable insights, is subject to several limitations. The selection of 50 articles focused on obesity and the use of a specific set of 15 STROBE-derived questions might not fully capture the breadth of medical research. Additionally, the reliance on binary and multiple-choice questions restricts the evaluation of LLMs' ability to provide nuanced answers. The rapid evolution of LLMs means that the findings might not be applicable to future versions, and potential biases within the training data have not been systematically assessed. Furthermore, the study's reliance on a single highly experienced medical professor as the benchmark, while valuable, might limit the generalizability of the findings. A larger panel of experts with diverse areas of specialization might provide a more comprehensive reference standard for evaluating LLM performance. Further investigation with a wider scope and more advanced methodologies is needed to fully understand the potential of LLMs in medical research.

In conclusion, LLMs show promise for transforming medical research, potentially enhancing research efficiency and evidence-based decision making. This study demonstrates that LLMs exhibit varying capabilities in understanding medical research articles. While newer models generally demonstrate better comprehension, no single LLM currently excels in all areas. This highlights the need for further research to understand the complex interplay of factors influencing LLM performance. Continued research is crucial to address these limitations and ensure the safe and effective integration of LLMs in healthcare, maximizing their benefits while mitigating risks.

Acknowledgments

We gratefully acknowledge Dr. Hilal Duzel for her invaluable assistance in validating the reference standard used in this study. Dr. Duzel's expertise in epidemiology and statistical analysis ensured the accuracy and robustness of the benchmark against which the LLMs were evaluated.

Conflicts of Interest

None declared.

Multimedia Appendix

Multimedia Appendix 1: [Percentages of Correct Answers by Large Language Models for Each Question]

REFERENCES

1. Lv Z. Generative artificial intelligence in the metaverse era. Cognitive Robotics. 2023;3:208-217. doi:10.1016/j.cogr.2023.06.001
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230-243. doi:10.1136/svn-2017-000101
3. Orrù G, Piarulli A, Conversano C, Gemignani A. Human-like problem-solving abilities in large language models using ChatGPT. Front Artif Intell. 2023;6. doi:10.3389/frai.2023.1199350
4. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online October 10, 2018. http://arxiv.org/abs/1810.04805
5. Chenais G, Gil-Jardiné C, Touchais H, et al. Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study. JMIR AI. 2023;2:e40843. doi: 10.2196/40843
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
7. Open AI. Accessed November 19, 2023. https://openai.com/blog/chatgpt/
8. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage. 2023;71:102642. doi:10.1016/j.ijinfomgt.2023.102642
9. Akyon SH, Akyon FC, Yılmaz TE. Artificial intelligence-supported web application design and development for reducing polypharmacy side effects and supporting rational drug use in geriatric patients. Front Med (Lausanne). 2023;10. doi:10.3389/fmed.2023.1029198
10. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
11. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-1131.e9. doi:10.1016/j.cell.2018.02.010
12. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. Journal of the American Academy of Orthopaedic Surgeons. 2023;31(23):1173-1179. doi:10.5435/JAAOS-D-23-00396
13. Preiksaitis C, Ashenburg N, Bunney G, et al. The Role of Large Language Models in

Transforming Emergency Medicine: Scoping Review. JMIR Med Inform. 2024;12:e53787. doi: 10.2196/53787
14. Kumar M, Mani UA, Tripathi P, Saalim M, Roy S. Artificial Hallucinations by Google Bard: Think Before You Leap. Cureus. Published online August 10, 2023. doi:10.7759/cureus.43313
15. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. J Med Internet Res. 2023;25:e46924. doi:10.2196/46924
16. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval Augmentation Reduces Hallucination in Conversation. ArXiv. Published online 2021. doi:https://doi.org/10.48550/arXiv.2104.07567
17. Cuschieri S. The STROBE guidelines. Saudi J Anaesth. 2019;13(5):31.
18. Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335(7624):806-808. doi:10.1136/bmj.39335.541782.AD
19. STROBE. STROBE Checklist: cohort, case-control, and cross-sectional studies (combined). Published 2023. Accessed December 28, 2023. https://www.strobe-statement.org/download/strobe-checklist-cohort-case-control-and-cross-sectional-studies-combined
20. Chen TJ. ChatGPT and other artificial intelligence applications speed up scientific writing. Journal of the Chinese Medical Association. 2023;86(4):351-353. doi:10.1097/JCMA.0000000000000900
21. Kitamura FC. ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment. Radiology. 2023;307(2). doi:10.1148/radiol.230171
22. Kumar M, Mani UA, Tripathi P, Saalim M, Roy S. Artificial Hallucinations by Google Bard: Think Before You Leap. Cureus. Published online August 10, 2023. doi:10.7759/cureus.43313
23. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11. doi:10.3389/fpubh.2023.1166120
24. Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V., & Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. Journal of medical internet research, 25, e51580
25. Wong RS-Y, Ming LC, Raja Ali RA. The Intersection of ChatGPT, Clinical Medicine, and Medical Education. JMIR Med Educ. 2023;9:e47274. doi: 10.2196/47274
26. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthcare. 2023;11(14):2046. doi:10.3390/healthcare11142046
27. Schubert MC, Lasotta M, Sahm F, Wick W, Venkataramani V. Evaluating the Multimodal Capabilities of Generative AI in Complex Clinical Diagnostics. Medrxiv. Published online 2023. doi:10.1101/2023.11.01.23297938
28. Eriksen A V., Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. NEJM AI. 2023;1(1). doi:10.1056/aip2300031
29. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. Journal of the American Academy of Orthopaedic Surgeons. 2023;31(23):1173-1179. doi:10.5435/JAAOS-D-23-00396
30. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health. 2023;11(4):e002391. doi:10.1136/fmch-2023-002391
31. Coşkun B, Bayrak O, Ocakoglu G, Acar HM, Kaygisiz O. Assessing the Accuracy of AI

Language Models in Providing Information on Urinary Incontinence: A Comparative Study. European Journal of Human Health. 2023;3(3):61-70. doi:10.29228/ejhh.71797
32. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus. Published online June 22, 2023. doi:10.7759/cureus.40822
33. King RC, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in answering heart failure related questions. 2023. doi:10.1101/2023.07.07.23292385
34. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clin Exp Dermatol. Published online June 2, 2023. doi:10.1093/ced/llad197
35. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. Published online March 20, 2023. http://arxiv.org/abs/2303.13375
36. Wu S, Koo M, Blum L, et al. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. NEJM AI. 2024;1(2). doi:10.1056/aidbp2300092
37. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv. Published online 2021. doi:10.48550/arXiv.2005.11401
38. Ferber, D., Wiest, I. C., Wölflein, G., Ebert, M. P., Beutel, G., Eckardt, J. N., ... & Kather, J. N. (2024). GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. NEJM AI, AIcs2300235. ,
39. Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics. 2023;39(9). doi:10.1093/bioinformatics/btad557

Abbreviations

Artificial intelligence (AI)

Machine learning (ML)

Natural language processing (NLP)

Large language models (LLMs)

ChatGPT (Chat Generative Pretrained Transformer)

GPT (Generative Pretrained Transformer)

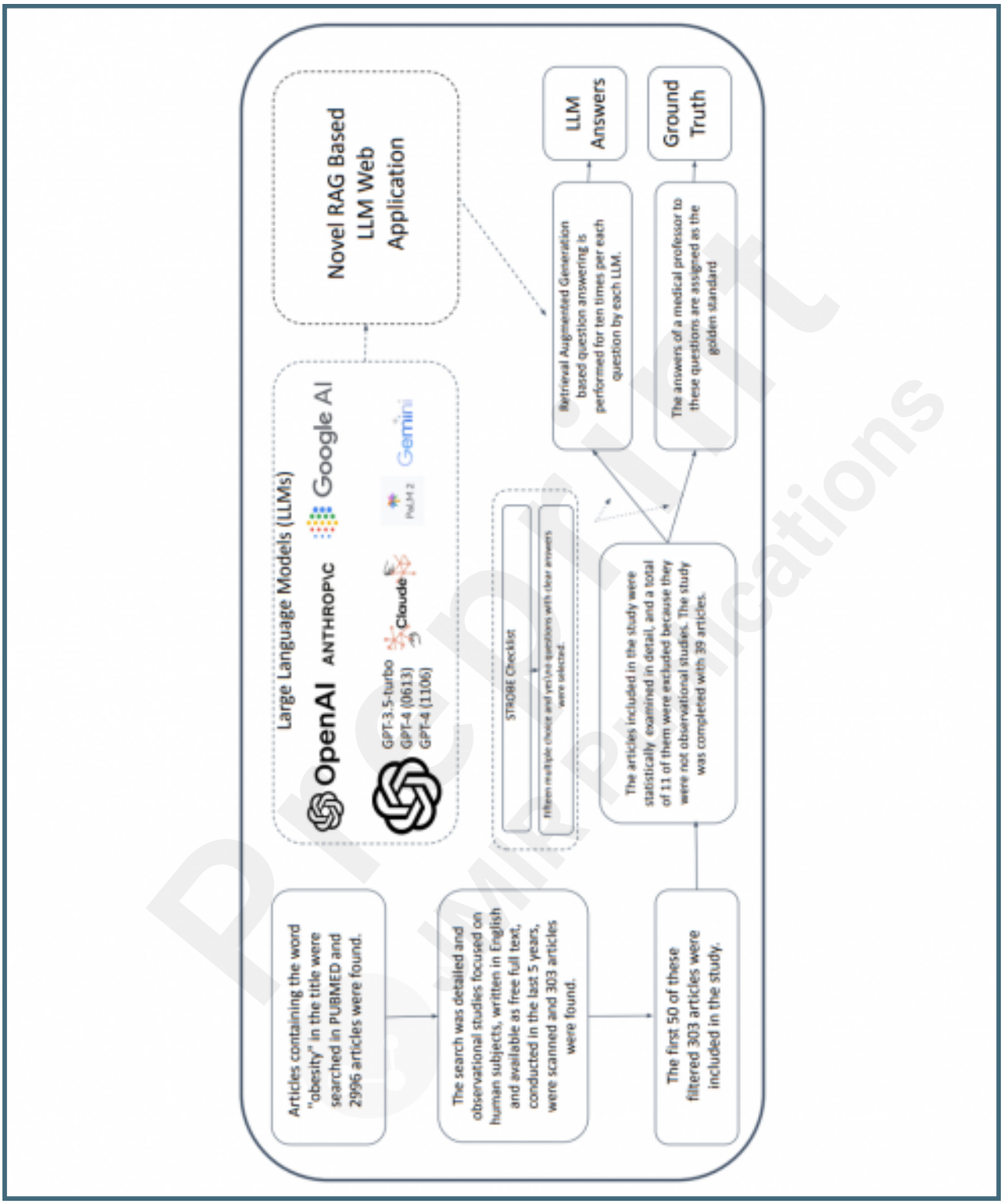
Retrieval augmented generation (RAG)

The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)

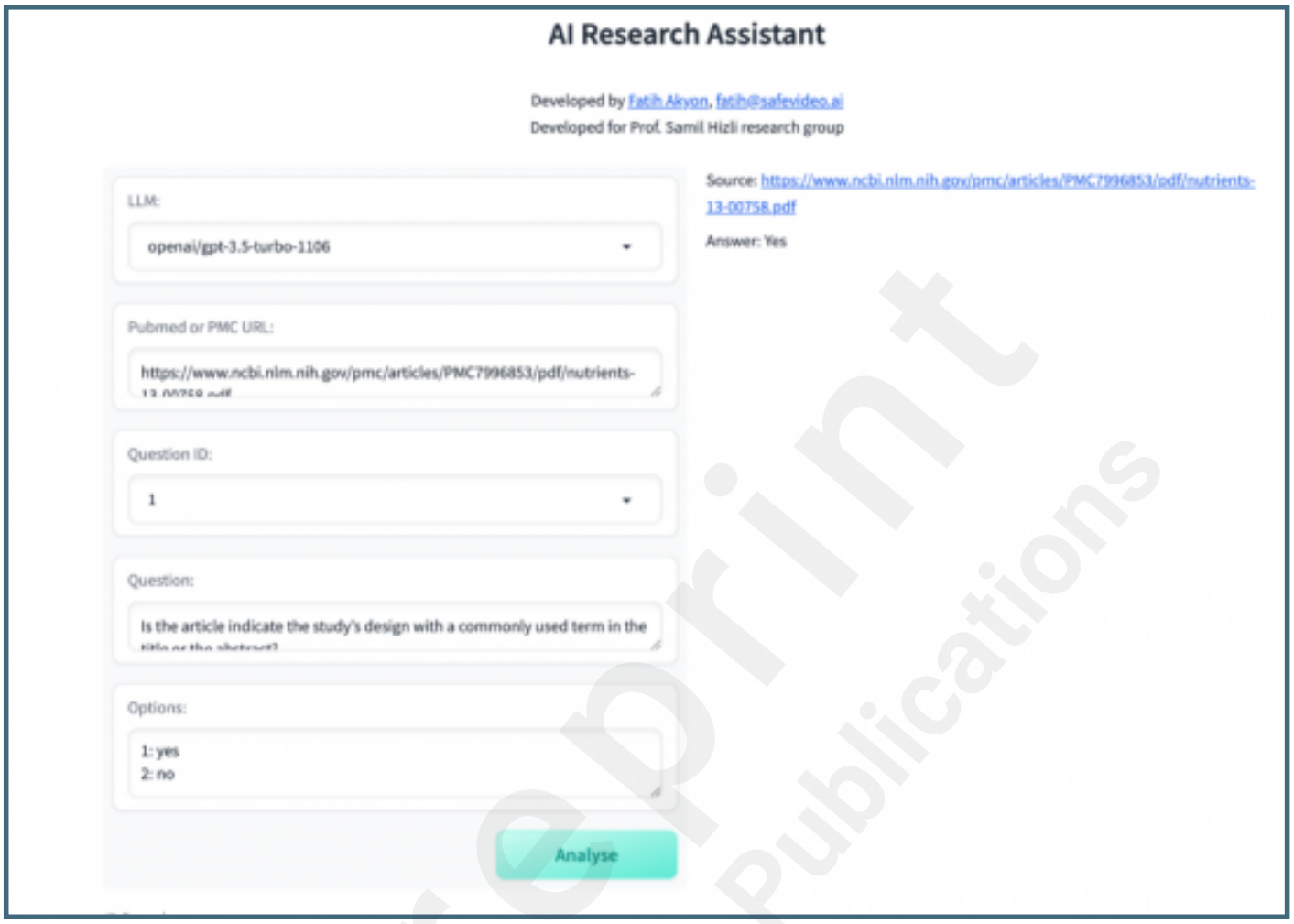
The United States Medical Licensing Exam (USMLE)

Supplementary Files

Untitled.

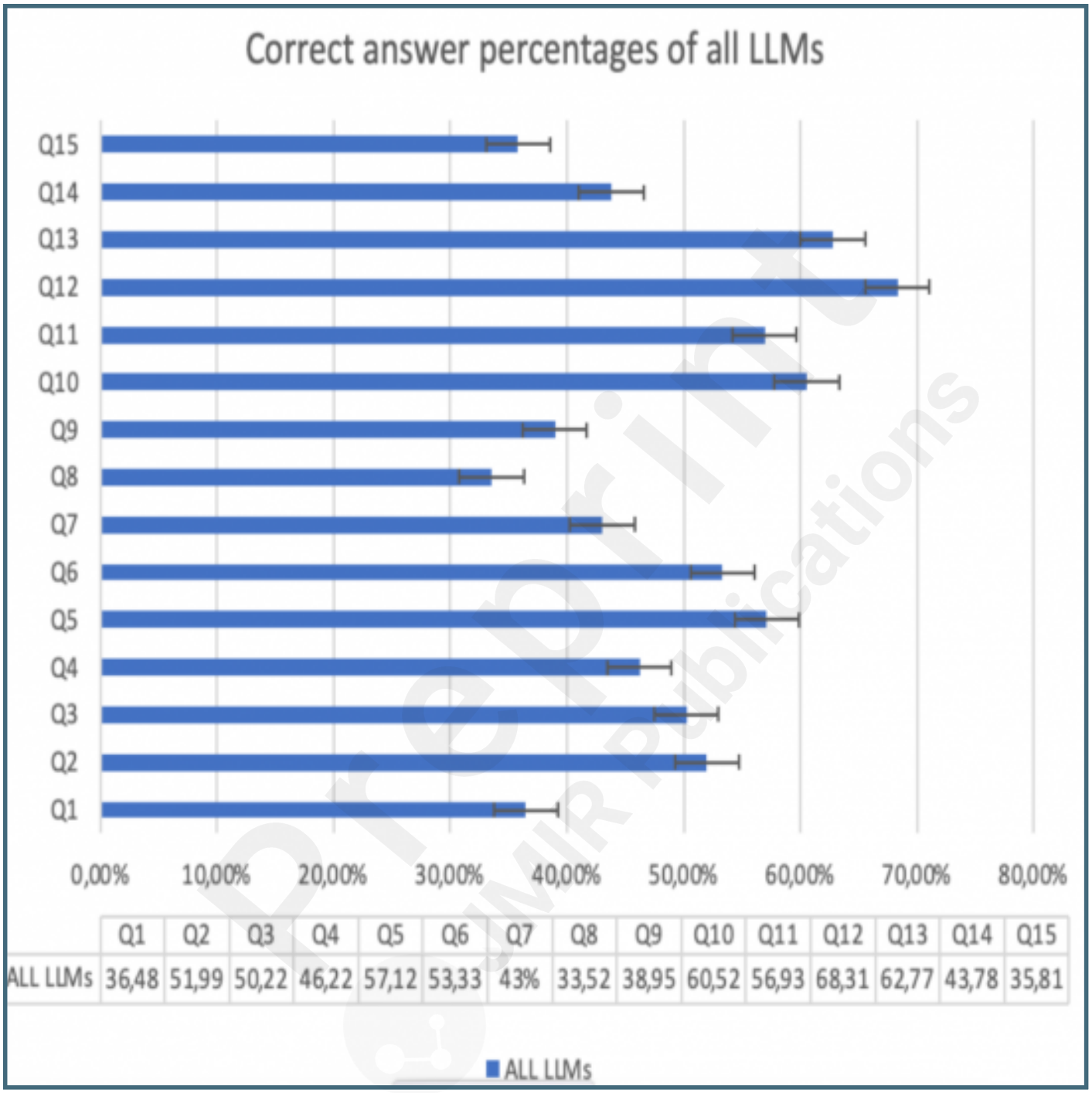

Novel RAG Based LLM Web Application
Large Language Models (LLMs)
OpenAI
ANTHROPIC
Google AI
Gemini
PaLM 2
Claude
GPT-3.5-turbo
GPT-4 (0613)
GPT-4 (1106)
LLM Answers
Ground Truth
Retrieval Augmented Generation based question answering is performed for ten times per each question by each LLM.
The answers of a medical professor to these questions are assigned as the golden standard
STROBE Checklist
Articles containing the word "obesity" in the title were searched in PUBMED and 2996 articles were found.
The search was detailed and observational studies focused on human subjects, written in English and available as free full text, conducted in the last 5 years, were scanned and 303 articles were found.
The first 50 of these filtered 303 articles were included in the study.
The articles included in the study were statistically examined in detail, and a total of 11 of them were excluded because they were not observational studies. The study was completed with 39 articles.

Untitled.



Figures

revised-Figure 3. Correct answer percentages of Large Language Models .png.



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
ALL LLMs	36,48	51,99	50,22	46,22	57,12	53,33	43%	33,52	38,95	60,52	56,93	68,31	62,77	43,78	35,81

Multimedia Appendixes

Percentages of Correct Answers by Large Language Models for Each Question.
URL: http://asset.jmir.pub/assets/7c70be706ca32e4f5f50c390a1aaecca.png