

Assessing language model performance in biomedical question-answering: A case study using the langchain framework on the CliCR Dataset

Feras ALMANNAA^{1*}

Ferdi SÖNMEZ²

Geliş tarihi / Received: 08.08.2025

Düzeltilerek geliş tarihi / Received in revised form: 13.08.2025

Kabul tarihi / Accepted: 13.08.2025

DOI: 10.17932/IAU.ABMYOD.2006.005/abmyod_v20i72005

Abstract

This paper focuses on developing and implementing a biomedical question-answering (BQA) system using large language models (LLMs) and the CliCR dataset, in combination with the LangChain framework. The study evaluates several models, including GPT-3.5, GPT-4, LLAMA3, and Mistral, in handling clinical questions. Key methodologies include data preparation, prompt engineering, and model adaptation. The evaluation employs metrics such as precision, recall, F1-score, BLEU scores, and embedding-based metrics. Results show that using the entire case context significantly outperforms chunking and vector store indexing methods. Notably, GPT-4 achieved an exact match score of 44.7%, surpassing human experts. Although fine-tuning improves domain-specific performance, there's a risk of overfitting. This research adds to the progress in BQA systems with possible benefits for clinical decision-making and medical education.

Keywords: Biomedical Question-Answering; CliCR; Evaluation; Large Language Models; LangChain; Prompt Engineering; RAG; Vector Database; Chunking

¹ Department of Artificial Intelligence and Data Science, Istanbul Aydin University, Istanbul Turkey Email : ferasmhddaymanalmanna@stu.aydin.edu.tr; ORCID: 0009-0005-4645-1071

² Department of Artificial Intelligence and Data Science, Istanbul Aydin University, Istanbul Turkey Email : ferdisonmez@aydin.edu.tr; ORCID: 0000-0002-5761-3867

Biyomedikal soru-cevaplamada dil modeli performansının değerlendirilmesi: CliCR veri seti üzerinde langchain çatısı kullanılarak bir vaka çalışması

Özet

Bu makale, LangChain çatısı ile birlikte büyük dil modellerini (BDM'ler) ve CliCR veri setini kullanarak bir biyomedikal soru-cevaplama (BSO) sistemi geliştirmeye ve uygulamaya odaklanmaktadır. Çalışma, GPT-3.5, GPT-4, LLAMA3 ve Mistral dahil olmak üzere çeşitli modellerin klinik soruları ele alma performansını değerlendirmektedir. Temel metodolojiler arasında veri hazırlama, komut mühendisliği (prompt engineering) ve model adaptasyonu yer almaktadır. Değerlendirmede kesinlik (precision), duyarlılık (recall), F1 skoru, BLEU skorları ve gömme (embedding) tabanlı metrikler gibi ölçütler kullanılmaktadır. Sonuçlar; vaka bağlamının tamamını kullanmanın, parçalara ayırma (chunking) ve vektör deposu indeksleme yöntemlerine göre önemli ölçüde daha iyi performans gösterdiğini ortaya koymaktadır. Dikkat çekici bir şekilde, GPT-4, %44,7'lik bir tam eşleşme skoru elde ederek insan uzmanları geride bırakmıştır. İnce ayar (fine-tuning) alana özgü performansı artırır da, aşırı öğrenme (overfitting) riski taşımaktadır. Bu araştırma, klinik karar verme ve tıp eğitimi için potansiyel faydalar sunarak BSO sistemlerindeki ilerlemelere katkıda bulunmaktadır.

Anahtar Kelimeler: *Biyomedikal Soru-Cevap; CliCR; Değerlendirme; Büyük Dil Modelleri; İstem mühendisliği, RAG; Vektör Veritabanı; Parçalama*

Introduction

The developments in large language models (LLM), such as GPT-3.5, GPT-4 (Achiam et al., 2024), and LLAMA (Touvron et al., 2023), have considerably affected the domains of natural language processing (NLP) and artificial intelligence (AI). These models can comprehend and produce human language, rendering them useful for biomedical applications. Almanac (Hiesinger et al., 2023) exhibited the possible biomedical uses of LLMs in question-answering tasks, assisting clinicians, researchers, and patients in procuring credible and timely information. BQA systems are proving to be a promising potential in answering complex clinical questions automatically, thus assisting in clinical decision-making, medical education, and research (Jin et al., 2021).

But, despite the presented potential, AMMU (Kalyan et al., 2021) shed light on some of the challenges that remained unsolved. A primary issue with BQA is the understanding and reasoning over the medical contexts, which involve technical terminology and complex relationships among medical entities. Current language models may struggle with these nuances and vocabularies, leading to inaccurate or incomplete answers (Jin et al., 2021). It is shown that fine tuning LLMs on extract domain-specific datasets improves the model's performance, but it has the disadvantage of easily overfitting and losing generalizability in low-resource domains (Gu et al., 2020), which remains a challenge. Duong and Solomon (2023) found that although GPT had nearly human-like performance on the memorization questions, it scored poorly on critical thinking questions. Lee et al. (2023) crafted a privacy-preserving LLM for automated data extraction with very good concordance to human reviewer gold standard, demonstrating both the potential and limitations of LLMs in biomedical contexts.

To address these challenges, our study explores the development of an advanced BQA system utilizing the latest developments of LLMs on the CliCR dataset (Suster & Daelemans, 2018), leveraging the LangChain framework . The methodology employed involves extensive data preparation, prompt engineering, and model adaptation using LangChain's APIs and tools, followed by measuring the results using multiple carefully designed evaluation metrics including precision, recall, F1-score, BLEU scores, and embedding-based metrics. The advantage of this approach lies in exploring and measuring different prompting techniques that both utilize full context windows and supporting vector stores, in combination with

sophisticated prompt engineering practices, enhancing the understanding and generation of accurate and relevant answers to biomedical queries. This research aims to advance the integration of LLMs in BQA systems, contributing to improved clinical decision-making and medical education.

To address these challenges, this study makes several key contributions:

- **Adoption of State-of-the-Art LLMs:** Using frontier large language models (LLMs), like GPT-3.5, GPT-4, LLAMA3, and Mistral, this study aims to enhance biomedical question answering (BQA) systems. By incorporating these advanced models, we hope to improve the performance of the BQA system.
- **Comprehensive Evaluation Across Multiple Models:** Performance comparison of closed, open, and fine-tuned LLMs on CliCR dataset which will help us emphasize strengths and limitations of models to boost the accuracy and reliability of BQA systems.
- **Prompt Engineering and Context Utilization:** Explore different prompt engineering and context retrieval techniques when evaluating LLMs performance within the BQA field, which builds upon the work of evaluating different approaches for applying LLMs in the biomedical field.
- **Fine-tuning and Domain Adaptation:** Finally, explore whether fine-tuning LLMs on domain specific data will improve the understanding and QA skills when dealing with medical cases. Previous work has showcased LLMs performance improvements in the biomedical field when fine-tuning on biomedical datasets (Gu et al., 2020; Tinn et al., 2021).

Through these contributions, the study aims to test, evaluate, and experiment with different technologies and techniques to help advance biomedical question-answering systems using the latest LLMs advancements. The work has been conducted through building a modular framework for efficient experimentations and evaluation of different prompts and retrieval techniques.

Literature review

The usage of LLMs in the biomedical space has generated considerable research interest. During our work, we reviewed recent studies and explored the performance and challenges of utilizing LLMs in the biomedical domain, while also looking at future insights.

-Performance evaluation

A case study has been run to assess the performance of LLMs in varying biomedical tasks. In genetics question answering, GPT was compared against a human to find that GPT was able to match human performance for factual recall but did not perform as well on critical thinking questions (Duong & Solomon, 2023). An evaluation of privacy-preserving LLMs for the automated data extraction of thyroid cancer pathology reports demonstrated that the model performed at the level of human reviewers, suggesting LLMs could be deployed for clinical data processing while safeguarding patient information (Lee et al., 2023). LLMs were tested against a neurologic board- style exam wherein the newer versions scored better than the average human, especially in behavioral, cognitive, and psychiatric domains (Schubert et al., 2023). Comparative scurrying of the ChatGPT-3.5, Google Bard, and Microsoft Bing on hematology cases found ChatGPT able to outperform the others in managing complex medical queries (Kumari et al., 2023).

-Challenges and limitations

LLMs face several challenges in biomedical applications. One study identified issues such as inconsistency and inaccuracy in LLM responses to pathology board exams, emphasizing the need for ongoing model refinement and human oversight (Koga, 2023). It was also found that GPT's performance in clinical diagnosis declined without detailed narrative context, highlighting the importance of comprehensive input for accurate diagnostic reasoning (Reese et al., 2023).

Furthermore, bias amplification has been revealed in LLMs, particularly in clinical phenotyping, where models tended to underdiagnose certain subpopulations, raising concerns about equitable healthcare applications (Pal et al., 2023).

-Innovative Applications

LLMs have shown potential beyond traditional question-answering tasks. An evaluation of GPT-4V on challenging medical cases involving both text and images showed that the model outperformed human respondents, suggesting multimodal AI could enhance medical diagnostic reasoning (Buckley et al., 2023). In biostatistics, ChatGPT demonstrated an ability to guide analyses of NHANES data, making complex statistical methods more accessible to non-experts (Titus, 2023).

LLMs have also been assessed in radiation oncology physics, where ChatGPT-4 performed on par with medical physicists, indicating a potential for these models to act as knowledgeable assistants in specialized fields (Holmes et al., 2023).

-Datasets and Methodologies

Developing sophisticated BQA systems depends on a knowledge of the current research landscape, which includes studying datasets and evaluation techniques. CliCR is a dataset introduced for machine comprehension in the medical domain, based on clinical case reports and consisting of about 100,000 gap-filling questions (Suster & Daelemans, 2018). This survey of BQA looks at the different approaches and challenges, noting that there is little real-world usage due to its immaturity (Jin et al., 2021). The other survey was on transformer-based biomedical pretrained language models (BPLMs) addressing their pre-training, tasks, and fine-tuning (Kalyan et al., 2021). Another survey looks into the advances in task-related domain-specific datasets, placing emphasis on the improvement of the accuracy of biomedical QA systems (Wang, 2022).

-Privacy and Ethical Considerations

Using LLMs in biomedical scenarios therefore raises privacy and ethical issues. The LLMs can, however, be used in clinical settings provided thorough anonymization of records is ensured so as not to compromise patient privacy (Lee et al., 2023). They have also looked at employing LLMs to answer routine patient questions regarding post-operative care, emphasizing the need for safety mechanisms and ethical guidelines to support responsible use in healthcare (Chowdhury et al., 2023).

-Future Directions

Integrating LLM-powered autonomous agents in simulated environments offers a new way to improve the learning and adaptability of medical agents. The simulation of hospital environments allows autonomous agents to learn and practice medical procedures, accumulating experience and refining their decision-making in a risk-free setting. Future work should look into developing such simulacra to provide continuous learning, potentially utilizing evolutionary learning techniques to create more robust and accurate BQA systems (Li et al., 2024).

Table 1. Comparative analysis of previous works.

Study	Objective	Model Used	Key Findings	Technical Depth
Duong and Solomon (2023)	Evaluate ChatGPT vs. human respondents on genetics questions	ChatGPT	ChatGPT performed well on memorization but struggled with critical thinking questions	Comparison with human performance, identification of strengths and weaknesses in critical reasoning
Lee et al. (2023)	Develop privacy-preserving LLM for data extraction from pathology	Custom LLM	High concordance rates with human reviewers, maintained patient privacy	Custom LLM development, focus on privacy-preserving techniques, application in clinical data extraction
Schubert et al. (2023)	Assess LLMs on neurology board-style examinations	Various LLMs	Newer LLMs outperformed human averages, particularly in behavioral and cognitive questions	Evaluation across multiple LLMs, detailed analysis of performance in specific medical fields
Kumari et al. (2023)	Compare ChatGPT-3.5, Google Bard, and Microsoft Bing on hematology	ChatGPT-3.5, Google Gemini, Bing	ChatGPT outperformed other models in handling complex medical queries	Comparative analysis of multiple models with focus on complex medical cases
Koga (2023)	Explore pitfalls of LLMs in pathology board examinations	GPT-4	Identified issues like inconsistency and inaccuracy in responses	Identification of LLM limitations, emphasis on the need for model refinement and human oversight
Reese et al. (2023)	Evaluate limitations of GPT-4 in clinical diagnosis	GPT-4	Performance dropped without detailed narrative context	Analysis of context importance in clinical diagnosis, evaluation of model limitations
Pal et al. (2023)	Study bias amplification in LLMs in clinical phenotyping	Various LLMs	LLMs underdiagnosed certain subpopulations	Examination of bias in LLMs, impact on equitable healthcare applications

Buckley et al. (2023)	Evaluate GPT-4V on text and image-based medical cases	GPT-4V	GPT-4V outperformed humans in multimodal medical diagnostic reasoning	Integration of text and image data, multimodal evaluation, performance comparison with humans
Titus (2023)	Use ChatGPT in biostatistics for analyzing NHANES data	ChatGPT	Simplified complex data analysis, made statistical methods accessible to non-experts	Application in biostatistics, focus on accessibility and simplification of complex analyses
Holmes et al. (2023)	Assess LLMs in radiation oncology physics	ChatGPT-4	ChatGPT-4 performed comparably to medical physicists	Evaluation in a specialized medical field, comparison with professional performance
Suster and Daelemans (2018)	Introduce CliCR dataset for machine comprehension in medical domain	Custom dataset	Dataset comprises 100,000 gap-filling queries from clinical case reports	Creation of a domain-specific dataset, focus on machine comprehension in the medical field
Jin et al. (2021)	Survey of BQA approaches and challenges	Various approaches	Identified underutilization in real-life settings due to system immaturity	Comprehensive survey, categorization of BQA approaches, identification of practical challenges
Kalyan et al. (2021)	Survey of transformer-based biomedical pretrained language models	Various BPLMs	Highlighted significance of BPLMs in NLP for biomedical domain	Detailed survey of transformer-based models, discussion of pre training methods, tasks, and fine-tuning techniques
Wang et al. (2022)	Understand domain-specific datasets for BQA	Various datasets	Emphasized the importance of specialized data for improving accuracy and relevance	Analysis of domain specific datasets, focus on the impact of specialized data on BQA performance

Li et al. (2024)	Explore use of autonomous agents in simulated hospital environments	Autonomous agents, LLMs	Agents learned and practiced medical procedures, improving decision-making processes	Simulation of hospital environments for agent training, evolutionary learning approaches in medical decision making
------------------	---	-------------------------	--	---

Methodology, materials and methods

We use recent large language models and NLP tools with the LangChain framework to build a Biomedical Question Answering evaluation system. The process is broken down into 3 stages: data, model and evaluation. Figure 1 provides a comprehensive overview of this process, with key steps outlined:

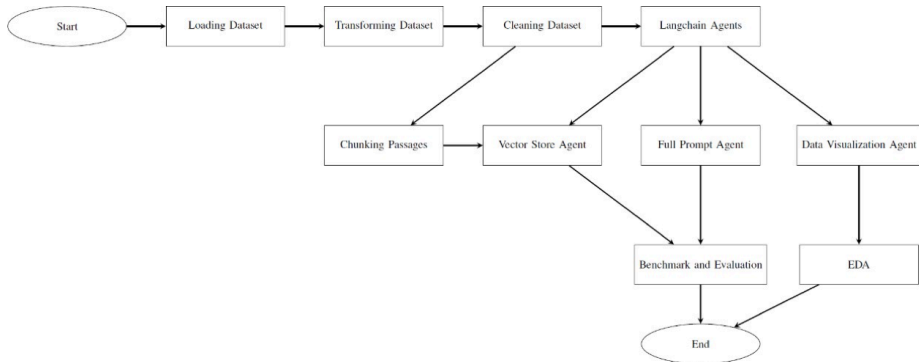


Figure 1. LangChain agents pipeline

- Data Preparation

In the initial phase, CliCR dataset has been processed to ensure compatibility with various large language models. The main steps include:

- A. Data transformation: This involves tasks such as loading the dataset, restructuring it, cleaning, vectorizing, and performing various text processing tasks. These tasks vary based on the chosen prompting strategy (discussed in the next step) and the specific model being adapted.
- B. Handling Long Passages: With an average passage length of approximately 1,466 tokens, the dataset fits within the context windows of most large language models. However, some passages exceed the

maximum context window of many local LLMs running on constrained environments, necessitating strategies to split the content into smaller, manageable chunks, potentially using a sliding window technique.

- C. Prompt engineering: The dataset is designed to impose restrictions on the predicted answer format. In this step, the most effective prompt likely to yield the correct answer has been engineered. Various versions of the prompt have been tested to align with the capabilities of the selected LLM.

- Model Adaptation with LangChain

Since the advancement of local and open LLMs has skyrocketed in the past few months, both the famous GPT model, which is a state-of-the-art closed large language model including its different versions, and state of the art open and local models like LLAMA3, Mistral, and Command-R have been employed to compare their performance. They have been adapted to the dataset by exploring multiple methods and prompts. To ensure the work remains LLM agnostic, an open-source generative language framework called LangChain has been utilized. LangChain provides a user- friendly interface for integrating LLM functionalities, enabling effortless switching between different LLMs, data storage solutions, and transformation tools. This flexibility allows the same processes to be applied across various LLMs and storage engines without significant modifications.

-Evaluation

The system's performance is assessed using various evaluation methods based on the original dataset creator's work. These benchmarks include exact match (EM), F1 score, BLEU scores (B-2 and B-4), and an embedding-based metric (E-avg) Table 2. We compare the results with the findings mentioned in the original dataset paper and also indicate future directions for our work.

-Large language models

This study evaluates the following LLMs:

- GPT (Generative Pre-trained Transformer): GPT-4 by OpenAI has a transformer architecture and was trained on many datasets and refined with reinforcement learning from human feedback (RLHF). It shows superior performance over GPT-3.5 on various standardized exams and in natural language understanding and generation (Achiam et al., 2024)

- **LLAMA3:** The LLaMA models from MetaAI range from 7B to 65B parameters. They use RMS normalization, SwiGLU activation functions, rotary positional embeddings, and an AdamW optimizer (Touvron et al., 2023). The original LLaMA has fewer parameters than GPT-3 and yet beats it on multiple benchmarks.
- **Mistral:** Mistral 7B is a high-efficiency 7-billion-parameter model that incorporates grouped-query attention and sliding window attention. It is optimized for reduced memory usage and enhanced speed, outperforming LLaMA 2 13B in benchmarks such as reasoning, math, and code generation (Jiang et al., 2023).
- **Cohere's Aya:** This is a multilingual, instruction-finetuned model covering 101 languages. Built on mT5 with 13B parameters, it was trained using the xP3x and Aya Collection datasets and is evaluated with a comprehensive suite for generative and discriminative tasks (Üstün et al., 2024).

Table 2. Description of evaluation metrics

Metric	Description	Equation
Exact Match (EM)	Measures the percentage of instances where the predicted answer exactly matches the ground truth answer.	$EM = \frac{1}{N} \sum_{i=1}^N I(A_i = P_i)$ <p>where A_i is the actual answer, P_i is the predicted answer, and I is the indicator function.</p>
F1 Score	Captures the overlap between the predicted and ground truth answers. It is the harmonic mean of precision and recall.	$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ $\text{Precision} = \frac{TP}{TP + FP}$ $\text{Recall} = \frac{TP}{TP + FN}$ <p>TP = True Positives, FP = False Positives, FN = False Negatives.</p>
BLEU-2	Assesses token contiguity up to 2-grams. It is a popular metric for machine translation that measures the similarity between the predicted sequence and the ground truth.	$BLEU = BP \cdot \exp \left(\sum_{n=1}^2 w_n \log p_n \right)$ <p>For BLEU-2, $N = 2$, where p_n is the precision of n-grams and w_n are the weights.</p> $BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{c}{r})} & \text{if } c \leq r \end{cases}$ <p>c is the length of the candidate translation and r is the length of the reference translation.</p>
BLEU-4	Similar to BLEU-2, but assesses token contiguity up to 4-grams.	$BLEU = BP \cdot \exp \left(\sum_{n=1}^4 w_n \log p_n \right)$ <p>For BLEU-4, $N = 4$. The brevity penalty BP is calculated as in BLEU-2.</p>

Embedding-based (E-avg) Score	Measures semantic relatedness between the predicted and ground truth answers by comparing their vector embeddings.	$E\text{-avg} = \frac{1}{N} \sum_{i=1}^N \cos(E_{A_i}, E_{P_i})$ <p>where E_{A_i} is the embedding of the actual answer, E_{P_i} is the embedding of the predicted answer, and \cos is the cosine similarity.</p>
-------------------------------	--	--

The integration of these LLMs within LangChain framework enables the development and testing of reusable BQA systems capable of providing accurate and reliable information for clinical queries while allowing easy model swapping when necessary.

The following models' weights have been chosen to evaluate:

- Closed - GPT-3.5
- Closed - GPT-4o
- Open - LLAMA3 7B
- Open - Mistral 7B
- Open - Cohere's Aya 8B

All open models have been evaluated using 4-bit quantization and were running locally on a GeForce 3070-ti GPU. The work outlined the possibility of having these locally running, small, and powerful LLMs running locally on consumer hardware without the need for deploying these models in centralized infrastructure, which is both a more accessible and privacy compliant approach.

-CliCR dataset overview

CliCR dataset (Suster & Daelemans, 2018) was specially created for the medical machine reading comprehension task. Derived from about 12,000 BMJ Case Reports published between 2005 and 2016, it contains some 100,000 gap-filling queries. What makes this dataset special is its methodology of creation: medical entities appearing within the "Learning points" section of each report are masked, and these entities are considered as ground-truth answers to the queries.

The CliCR dataset features make it a strong benchmark. The passages themselves are varied and rich with linguistic details; providing on an average about 1,500 tokens of context for analysis. The answers consist mostly of medical terms that are multiword answer phrases, ranging from single words to long phrases, testing a model's proficiency in generating exact answers. Although this puts a challenge, especially when considering certain LLMs and their context window limitations, it serves as an excellent

dataset for evaluating and fine-tuning the models for genuine medical use cases.

Exploratory data analysis

In order to design methods best suited for this data, we carried an exploratory data analysis on the CliCR dataset. The main insights that we discovered during the analysis are going to contribute to the design of how we prompt LLMs during question answering, mainly:

- A. Prompt Engineering: By understanding the variations in queries, we can design prompts that elicit accurate responses from our models.
- B. Retrieval augmented generation (RAG): When we analyse the connections between search queries and specific sections of text, it helps us come up with methods to find the most relevant information.

We start by listing the key statistics in Table 3. This includes the total number of cases (11,846) and tokens in the passages (16,544,217 without recounting and 153,784,539 with recounting) so we can have an idea about the overall shape of the dataset.

- Answer Length Distribution: Figure 2 shows that most answers are between 1 to 5 words, 2 words being the most common. The dataset also has longer answers, some more than 20 words. So, we need to be able to instruct the models to generate short and limited answers without over-explaining, which is what they tend to do.
- Expression Variability: 56,093 unique answers, 288,211 with variations in expression. Which means we need to be able to generalize across different expressions of the same concept and have a rich vocabulary support.
- Common Answer Terms: Common terms like "treatment", "symptoms", "surgery", "MRI" in the case reports. This helps us define what medical knowledge the model needs to learn and apply and might help us when exploring medically fine-tuned LLMs.
- Answer Embedding in Passages: 58.94% of answers are embedded in the passages. That is a very important insight because it indicates that the LLM will need to understand the context to generate accurate answers in around 40% of the queries.
- Answer Retrieval: 88.67% of answers can be found in any passage in

the dataset. This affects how the LLM needs to generate new answers and requires it to be able to generate novel answers.

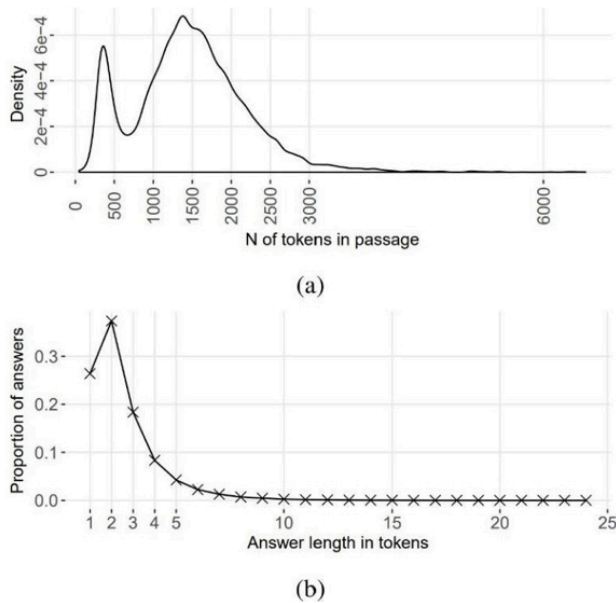


Figure 2. Distribution of answer lengths in the CliCR dataset

Table 3. Statistical overview of the CliCR dataset.

Data Statistics	
Number of cases	11,846
Number of queries	104,919
Total token count in passages without recounting	16,544,217
Total token count in passages with recounting	153,784,539
Average length of passages (tokens)	1,465.75
Number of unique word types in passages	112,673
Number of unique entity types in passages	591,960
Queries in the training set	91,344
Queries in the development set	6,391
Queries in the test set	7,184
Answer type: problem	%71.92
Answer type: treatment	%17.64
Answer type: test	%10.43

Results

We present our results after testing out different approaches on our question-answering system. We've explored prompt engineering, different LLMs, vector store indexing and fine-tuned models. We selected the best performing LLMs to date, both open and proprietary, to do a full comparison. We measured their performance, time and cost to see what each model can do and what the limitations are. This helped us figure out what's top performing during different setups.

To save on the compute resources, we limited the testing dataset to one question per case. This reduced inference time and cost while still giving us representative results, a balance between efficiency and accuracy.

First, we compared using a full case in the prompt with vector store indexing against the same LLM. The vector store indexing approach retrieved relevant case context chunks based on semantic similarity. We broke the dataset into smaller semantically related chunks and indexed them to fetch the most relevant chunks for each query during runtime. But this approach underperformed compared to the full prompt engineering approach, showing its limitation for this setup.

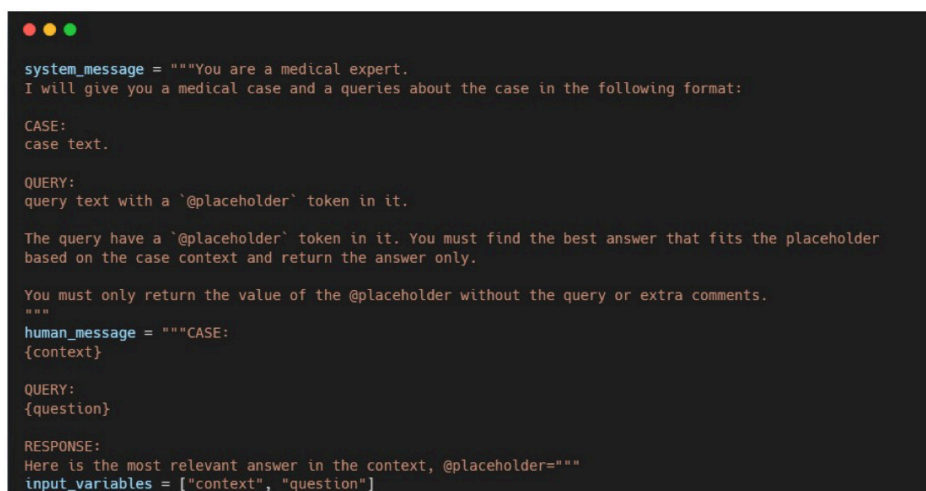
Table 4 shows us the results of the two approaches and how the full case approach performed better on all evaluation metrics. These results match up with the insights we got from the EDA step, when we identified that less than 60% of the answers are embedded within the case context. This led us to the conclusion that the LLMs will need more context to properly understand and generate the correct answer for the medical query.

Table 4. GPT3.5 full case/vector retrieval.

Method	EM	F1	BLEU-2	BLEU-4	E-avg
GPT-3.5 - Full Case	36.9%	53.1%	0.35	0.17	0.76
GPT-3.5 - Retrieval	23.9%	36.9%	0.19	0.07	0.68

When exploring fine-tuning possibilities, we evaluated the top open biomedical LLM called OpenBioLLM (paper to be released) using our established workflow. This model produced the most hallucinations and largely ignored our crafted prompts. This outcome suggests a huge loss in instruction-following capability and language understanding, likely due to overfitting on the medical dataset. Due to its extremely poor initial results, we excluded this model from further comparisons.

During evaluation, multiple prompts have been crafted, tested, and evaluated for both full case and the vector store cases respectively. Figure 3 showcases the top performing prompt, which as mentioned before was a full case prompt. The system prompt gives the LLM a medical expert personality, presents an example of a case format, and gives instructions on how to parse the coming cases. Then each case will be a simply formatted case as per the format presented in the system message. The human message (the case being tested) ends with an '=' symbol to instruct the model to give the final answer without trying to further explain the answer. This was needed as LLMs naturally tend to over-explain their answers.



```
system_message = """You are a medical expert.
I will give you a medical case and a queries about the case in the following format:

CASE:
case text.

QUERY:
query text with a '@placeholder' token in it.

The query have a '@placeholder' token in it. You must find the best answer that fits the placeholder
based on the case context and return the answer only.

You must only return the value of the @placeholder without the query or extra comments.
"""
human_message = """CASE:
{context}

QUERY:
{question}

RESPONSE:
Here is the most relevant answer in the context, @placeholder="""
input_variables = ["context", "question"]
```

Figure 3. Top performing prompt.

To ensure a proper assessment of performance, we ported the same evaluation metrics from the original dataset authors' work, as detailed in Table 2, while laying the groundwork for implementing more evaluation metrics.

Our results show that while vector store indexing is a way to get relevant information, full prompt engineering performed better on all evaluation benchmarks. This proves context is key and context engineering is essential for Q/A in biomedicine.

These results are useful for anyone looking to improve Q/A in biomedicine. They show you need to consider the needed context when asking medical questions and that certain approaches like chunking or fine tuning on too specific datasets have limitations in this domain.

Table 5. Evaluation Results

Method	EM	F1	BLEU-2	BLEU-4	E-avg	%
Human baseline						
Human Expert	35%	53.7%	0.46	0.23	0.67	100%
Human Novice	31%	45.1%	0.43	0.24	0.62	92.5%
Previous Work						
rand-entity	1.4%	5.1%	0.03	0.01	0.23	34.3%
maxfreq- entity	8.5%	12.6%	0.10	0.05	0.31	46.2%
sim-entity	20.8%	29.4%	0.22	0.15	0.45	67.1%
lang-model	2.1%	3.5%	0.00	0.00	0.30	44.7%
SA-Anonym	19.6%	27.2%	0.22	0.16	0.43	64.1%
SA-Ent	6.1%	11.4%	0.07	0.05	0.31	46.2%
GA-Anonym	24.5%	33.2%	0.28	0.20	0.48	71.6%
GA-Ent	22.2%	30.2%	0.25	0.18	0.46	68.6%
GA-NoEnt	14.9%	33.9%	0.21	0.11	0.51	76.1%
Our Work						
GPT-4o	44.7%	60.9%	0.43	0.24	0.80	119.4%
GPT-3.5	36.9%	53.1%	0.35	0.17	0.76	113.4%
AYA	34.3%	50.6%	0.42	0.27	0.70	104.4%
Mistral	20.8%	35.0%	0.22	0.11	0.64	95.5%
LLAMA3	17.1%	31.3%	0.12	0.05	0.61	91%

Limitations

While this study has demonstrated the promising capabilities of LLMs in the biomedical domain, it must be noted that the test was not without limitations. Our evaluation comprised a limited set of prominent LLMs and only contrasted a full-context approach with a standard vector-search retrieval method. The development in the AI is very fast and it might well be that other models or more sophisticated retrieval techniques yield different results. Additionally, based on our preliminary experiments on fine-tuning a biomedical LLM, we deemed that there might be issues with overfitting and a loss of instruction-following capability. There should thus be further investigations into methods for fine-tuning that create a good balance between domain-specific knowledge and general reasoning ability.

Furthermore, transferring a generic solution from research and adapting it to a custom solution in a real clinical use may impose practical constraints that must always be considered such as:

-Cost and Privacy

Proprietary models, such as the beloved GPT-4o, offer the highest performance but run only on external servers, which makes API costs a concern for patient data privacy and operational budgets.

-Accessibility and Infrastructure

Open-source models like AYA and LLAMA3 stand as strong alternatives since they can be running on local hardware, hence ensuring privacy and eliminating prohibitive long-term costs. There is, however, an upfront capital investment in computational infrastructure, and one must also deal with the short context window of any current locally run model, which becomes a bottleneck for very long clinical reports with hundreds of pages.

Finally, performance gaps detailed in Table 5 are observable. Notably, GPT-4o even scored above the human expert baseline. But to move beyond mere comparison and formally verify those results, the execution of a statistical significance test marks the very next step. For example, by applying a McNemar's test or bootstrapping test one could find out if the difference in performances. But due to limitation in accessing the results reported by original work we note this limitation.

Conclusion

In this study, we have developed and implemented a BQA system to measure the performance of different LLMs on the CliCR dataset. Despite some limitations, we presented some key observations out of our research. By having an extensible, model-agnostic framework with LangChain, we compared proprietary models such as GPT-4o with open-source ones such as LLAMA3, Aya, and Mistral.

A primary conclusion is that full context of an entire clinical case outperforms naïve chunking-based retrieval methods immensely. It brings to light the fact that when it comes to complex medical reasoning, often where answers are not given by implication, the whole context must be provided to the model to be able to respond properly. The combo of GPT-4o, landing at an EM score of 44.7%, which also outperformed human experts-on-ceiling, constitutes an indication that such systems can work in clinical workflows. On the other hand, the local model, Aya, gave another signal of another viable path toward private and secure clinical applications by nearly matching expert accuracy in performance.

Future work

Evaluation of methods for fine-tuning remains an interesting avenue for further investigation. Since initial results suggest little advantage was achieved by finetuned models, there could be more losses suffered in their general skills while overly fitted on the dataset they were fine-tuned upon, and hence it warrants a more thorough investigation. Thus, a complete evaluation of the fine-tuning approach and strategies would offer sound insight into its usability for a QA system.

In addition, the world of simulated training environments form important insights and directions for the further study of BQA systems (Li et al., 2024). Armed with simulated training environments, evolutionary learning methods, and comprehensive knowledge integration, future BQA systems may be refined and enhanced for increased accurateness and usefulness to clinical decision-making and medical education.

Finally, while our study showed that full context is performing better than the naïve RAG implementation evaluated in this study, it is notable that new and more advanced RAG techniques are emerging future research should focus on exploring these techniques on this and other medical datasets with even larger context windows.

Acknowledgments

The authors thank to Istanbul Aydin University for research support for the preparation of this joint work.

References

- [1]Buckley, T. A., Diao, J. A., Rodman, A., & Manrai, A. K. (2023). Accuracy of a vision-language model on challenging medical cases. *ArXiv, abs/2311.05591*. <https://arxiv.org/abs/2311.05591>
- [2]Chowdhury, M., Lim, E., Higham, A., McKinnon, R., Ventoura, N., He, Y., & Pennington, N.D. (2023). Can large language models safely address patient questions following cataract surgery? *Clinical NLP, 1*, 131–137. <https://aclanthology.org/2023.clinicalnlp-1.17>
- [3]Duong, D., & Solomon, B. D. (2023). Analysis of large-language model versus human performance for genetics questions. *medRxiv: The Preprint Server for Health Sciences*. <https://www.medrxiv.org/content/10.1101/2023.01.27.23285115v1>

- [4]Gu, Y., Tinn, R., Cheng, H., Lucas, M. R., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3, 1–23. <https://arxiv.org/abs/2007.15779>
- [5]Hiesinger, W., Zakka, C., Chaurasia, A., Shad, R., Dalal, A. R., Kim, J., Moor, M., Alexander, K., Ashley, E. A., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., & Nelson, J. (2023). Almanac: Retrieval-augmented language models for clinical medicine. *Research Square*. <https://arxiv.org/abs/2303.01229>
- [6]Holmes, J., Liu, Z., Zhang, L.-C., Ding, Y., Sio, T., Mcgee, L., Ashman, J., Li, X., Liu, T., Shen, J., & Liu, W. (2023). Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology*, 13. <https://arxiv.org/abs/2304.01938>
- [7]Holmes, J., Peng, R., Li, Y., Hu, J., Liu, Z., Wu, Z., Zhao, H., Jiang, X., Liu, W., Wei, H., Zou, J., Liu, T., & Shao, Y. (2023). Evaluating multiple large language models in pediatric ophthalmology. *ArXiv, abs/2311.04368*. <https://arxiv.org/abs/2311.04368>
- [8]Holmes, J., Ye, S., Li, Y., Wu, S.-N., Liu, Z., Wu, Z., Hu, J., Zhao, H., Jiang, X., Liu, W., Wei, H., Zou, J., Liu, T., & Shao, Y. (2023). Evaluating large language models in ophthalmology. *ArXiv, abs/2311.04933*. <https://arxiv.org/abs/2311.04933>
- [9]Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Sciao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *ArXiv*. <https://arxiv.org/abs/2310.06825>
- [10] Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., & Yu, S. (2021). Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55, 1–36. <https://arxiv.org/abs/2102.05281>
- [11] Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). AMMU: A survey of transformer- based biomedical pretrained language models. *ArXiv, abs/2105.00827*. <https://arxiv.org/abs/2105.00827>
- [12] Koga, S. (2023). Exploring the pitfalls of large language models:

Inconsistency and inaccuracy in answering pathology board examination-style questions. *Pathology International*, 73. <https://onlinelibrary.wiley.com/doi/10.1111/pin.13382>

- [13] Kumari, A., Kumari, A., Singh, A., Singh, S., Juhi, A., Dhanvijay, A., Pinjar, M., & Mondal, H. (2023). Large language models in hematology case solving: A comparative study of chatgpt- 3.5, google-bard, and microsoft bing. *Cureus*, 15. <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.19738>
- [14] Lee, D. T., Vaid, A., Menon, B. M., Freeman, R. R., Matteson, D. S., Marin, M. P., & Nadkarni, G. N. (2023). Development of a privacy preserving large language model for automated data extraction from thyroid cancer pathology reports. *medRxiv*. <https://www.medrxiv.org/content/10.1101/2023.11.07.23298000v1>
- [15] Li, J., Wang, S., Zhang, M., Li, W., Lai, Y., & Kang, X. (2024). Agent hospital: A simulacrum of hospital with evolvable medical agents. *ArXiv*. <https://arxiv.org/abs/2405.02957>
- [16] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, S., & Altman, S. (2024). GPT-4 Technical Report. *arXiv Preprint arXiv:2303.08774*. <https://arxiv.org/abs/2303.08774>
- [17] Pal, R., Garg, H., Patel, S., & Sethi, T. (2023). Bias amplification in intersectional subpopulations for clinical phenotyping by large language models. *medRxiv*. <https://www.medrxiv.org/content/10.1101/2023.03.22.23287585v1>
- [18] Reese, J., Danis, D., Caulfield, J. H., Casiraghi, E., Valentini, G., Mungall, C., & Robinson, P. N. (2023). On the limitations of large language models in clinical diagnosis. *medRxiv*. <https://www.medrxiv.org/content/10.1101/2023.07.13.23292613v1>
- [19] Schubert, M., Wick, W., & Venkataramani, V. (2023). Performance of large language models on a neurology board-style examination. *JAMA Network Open*, 6. <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2812620>
- [20] Suster, S., & Daelemans, W. (2018). CliCR: a dataset of clinical case reports for machine reading comprehension. *ArXiv*, *abs/1803.09720*. <https://arxiv.org/abs/1803.09720>

- [21] Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Fine-tuning large neural language models for biomedical natural language processing. *Patterns*. <https://arxiv.org/abs/2112.07869>
- [22] Titus, A. J. (2023). NHANES-GPT: Large language models (LLMs) and the future of biostatistics. *medRxiv*. <https://www.medrxiv.org/content/10.1101/2023.12.13.23299830v1>
- [23] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, F., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *ArXiv*. <https://arxiv.org/abs/2302.13971>
- [24] Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., & Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model. *ArXiv*. <https://arxiv.org/abs/2402.07827>
- [25] Wang, Z. (2022). Modern question answering datasets and benchmarks: A survey. *ArXiv*, *abs/2206.15030*. <https://arxiv.org/abs/2206.15030>