Can GPT Redefine Medical Understanding? Evaluating GPT on Biomedical Machine Reading Comprehension

Shubham Vatsal, Ayush Singh
inQbator AI at eviCore Healthcare
Evernorth Health Services
firstname.lastname@evicore.com

Abstract

Large language models (LLMs) have shown remarkable performance on many tasks in different domains. However, their performance in contextual biomedical machine reading comprehension (MRC) has not been evaluated in depth. In this work, we evaluate GPT on four contextual biomedical MRC benchmarks. We experiment with different conventional prompting techniques as well as introduce our own novel prompting method. To solve some of the retrieval problems inherent to LLMs, we propose a prompting strategy named Implicit Retrieval Augmented Generation (RAG) that alleviates the need for using vector databases to retrieve important chunks in traditional RAG setups. Moreover, we report qualitative assessments on the natural language generation outputs from our approach. The results show that our new prompting technique is able to get the best performance in two out of four datasets and ranks second in rest of them. Experiments show that modern-day LLMs like GPT even in a zero-shot setting can outperform supervised models, leading to new state-of-the-art (SoTA) results on two of the benchmarks.

1 Introduction

Machine Reading Comprehension (MRC) is defined as a task where a system tries to answer a question based on a given context. The context could be anything ranging from a couple of passages to a list of documents. Even though much research has been conducted on MRC, several challenges remain when dealing with MRC tasks (Sugawara et al., 2022), such as the inability to handle long-range dependencies when trying to do reasoning and domain adaptation. Recent improvements in large language modeling has alleviated a lot of the aforementioned issues.

MRC in the biomedical domain (Hermann et al., 2015; Baradaran et al., 2022) has always been a key area of research. Solving a biomedical MRC task faces various challenges including large intricate in-domain vocabulary, dependency on global knowledge, etc. Due to these challenges, there is a wide gap between the performance of conventional methods in the general domain and that of the biomedical domain. Although, traditional machine learning models did show some improvement, they have never been any close to human baselines or gold standards. Contrary to this preconceived notion, modern-day LLMs have shown remarkable performance on many biomedical tasks (Nori et al., 2023; Yang et al., 2023; Cheng et al., 2023).

MRC can have different variations in itself. A contextual MRC requires the LLMs to answer a query solely by relying on a given context. In contrast, a context-free MRC relies on model's embedded knowledge base or any open-source knowledge base, such as Wikipedia, to answer a query instead of using only the context provided. Some of the datasets corresponding to context-free MRC are Zhang et al. (2018); Pal et al. (2022). These datasets are classified under context-free MRC because the given context is not sufficient to answer the questions. There is a definite need to explore the LLM's inherent knowledge base or any other source of knowledge to answer these queries. Similarly, Berant et al. (2014b); Pappas et al. (2018); Zhu et al. (2020) comprise of the datasets for contextual MRC. Again, these datasets have been categorized under contextual MRC because the queries asked can be correctly answered just by looking at the provided context. There is no need to induce any kind of external knowledge in order to answer these questions.

Recent LLMs have attained unprecedented performance in a wide array of natural language processing (NLP) tasks (Chang et al., 2023). Although their performance have been evaluated on a multitude of MRC benchmarks in a context-free setting, their performance in a contextual setting has been

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 256–265
August 16, 2024. ©2024 Association for Computational Linguistics

Dataset	ProcessBank	BioMRC	MASH-QA	CliCR
# QA Pairs	150	6250	3493	7184
Avg Context Length	85	255	863	1461
Max Context Length	266	510	2911	3952

Table 1: Corpus Level Statistics

understudied. In this work, we fill out this missing gap by evaluating GPT (OpenAI, 2023) on standard contextual MRC benchmarks of the biomedical domain. The key contributions are:

1. We evaluate different prompting techniques with GPT on four contextual biomedical MRC benchmarks and report new SoTA results.

2. We propose a novel prompting method *Implicit RAG*. In this method, the LLM is asked to first retrieve the sections or textual extracts from the context that may be relevant to the query and then answer the given query. This technique shows that unlike conventional RAG we no longer need vector databases to store the embeddings of the entire corpus. It further emphasizes that LLMs are capable enough to do the retrieval in one go. Experiments show that this technique is able to achieve the best results in two out of four discussed datasets and ranks second in rest of them.

3. Although machine evaluation is a good measure of performance, it falls short when evaluating artificially generated text (Schluter, 2017), where actual human preferences are significantly superior. Therefore, we report qualitative preference metrics by human experts on the output of our proposed approach *Implicit RAG*. We find that humans agree with the generated outputs most of the time.

2 Related Work

MRC evaluates a system's ability to comprehend and then reason to answer questions over the natural language present in a passage or context. Over the years, quite a few variations of this task have been devised to address and evaluate various aspects of a MRC system namely cloze-style (Hermann et al., 2015; Yagcioglu et al., 2018; Pappas et al., 2018, 2020), multiple-choice (Richardson et al., 2013; Lai et al., 2017; Berant et al., 2014a), extractive (Yang et al., 2015; Trischler et al., 2016; Zhu et al., 2020) and generative QA (Nguyen et al., 2016; Kočiskỳ et al., 2018). In this study, we strive to evaluate three of the above discussed forms of MRC in the biomedical domain which are cloze-style, extractive and multiple-choice using GPT.

In order to elicit an answer from an LLM like GPT, one needs to prompt it in natural language in an optimal manner to retrieve the intended answer. To that end, there has been tremendous development in finding optimal methods for prompting LLMs. The maximum performance boost has been seen from the Chain-of-Thought (CoT) Reasoning (Wei et al., 2022) prompting strategy which asks the LLM to explain how it arrived at the answer. More recently, Analogical Reasoning (AR) (Yasunaga et al., 2023a) has been proposed that achieves drastically better performance than CoT and other prompting techniques. AR works by asking the LLM to reason about a problem by giving analogies which in return forces the model to leverage the global knowledge encoded in it. While prompting methods like CoT and AR improve LLM's performance by exploiting the model's global knowledge embedded in it, there has been an increase in developing novel techniques, especially for cases where the context that needs to be searched through to answer the asked query is huge. The context could be one huge document or a combination of multiple short/long documents. In such scenarios, it is very important to identify only the relevant chunks of the context required for the underlying task and pay attention to them. These emerging methods come under the umbrella of Retrieval Augmented Generation (RAG) (Lewis et al., 2020), which has been shown to improve the performance of LLMs by retrieving contextually relevant information from corpora. The basic methodology behind RAG is to use, embed, and store context in a vector database. These embeddings can then be retrieved based on their semantic similarity to the query.

All the aforementioned prompting methods help interface and contextualize inputs in a better manner for LLMs. While the efficacy of these methods has been seen on several large benchmarks in different domains, however, the degree to which they help in contextual biomedical MRC has been understudied. Mahbub et al. (2022) presented an adversarial learning-based domain adaptation frame-

You are a *{profession}* who is given a *{context_type}* and a corresponding *{query_type}*. Your job is to read the given *{context_type}* and then select the best option from a list of options to answer the *{query_type}*.

The *{query_type}* that needs to be answered is listed below.

{query_type}: *{query_text}*
List of options: *{options}*

Here is the *{context_type}* that needs to be read to select the best option from a given list of options for the *{query_type}*.

{context_text}

Figure 1: Basic Prompt Template

work for the biomedical MRC task to address the discrepancies in the marginal distributions between the datasets of the general and biomedical domains. Nori et al. (2023) evaluated GPT on medical competency examinations and benchmark datasets. Even though their work talks about biomedical MRC, it concentrates only on context-free MRC benchmarks. Similarly, Singhal et al. (2023) evaluated Med-PaLM 2 on medical competency examinations and thus focuses on context-free biomedical MRC.

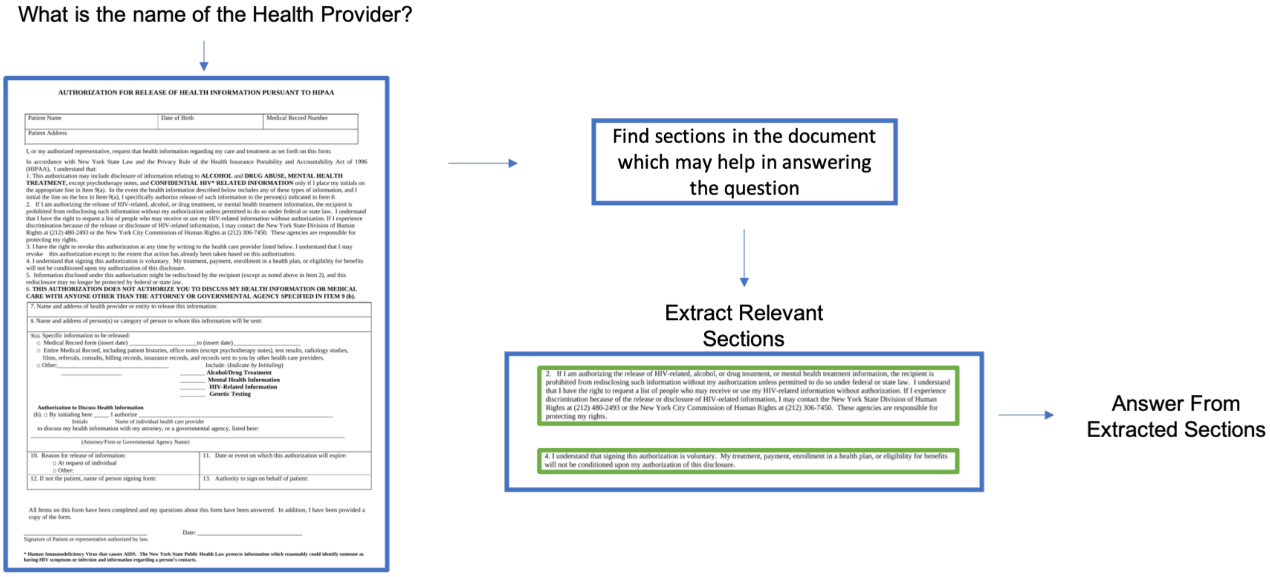
3 Datasets

The four datasets from the biomedical and healthcare domain we choose to explore and analyze the performance of GPT are ProcessBank (Berant et al., 2014b), BioMRC (Pappas et al., 2020), MASH-QA (Zhu et al., 2020) and CliCR (Šuster and Daelemans, 2018). There are a multiple reasons for selecting these four datasets. First, we want to focus on datasets that have not yet been evaluated by modern-day LLMs like GPT. Next, we want to pick up datasets that vary in their statistics and nature. Finally, based on our understanding, these 4 datasets covered the majority of research around contextual MRC in the biomedical field.

ProcessBank contains descriptions of biological processes as context accompanied by multiple-choice questions. BioMRC, an improved version of BioREAD (Pappas et al., 2018) is a large-scale cloze-style dataset. It contains abstracts and titles of biomedical articles and the task of any MRC system is to predict the missing entity in a title using the corresponding abstract as context. In BioMRC, all the biomedical entities mentioned in the abstract are considered as candidate answers and thus one option needs to be chosen from them. MASH-QA is associated with consumer health domain where the answers can consist of sentences from multiple spans of the long context. The candidate answers include every single sentence of the given context. CliCR is again a cloze-style dataset. It contains cloze queries from clinical case reports. Unlike BioMRC, CliCR doesn't really have a list of candidate answers with one of them being the correct answer. Rather, CliCR contains a ground-truth answer set which consists of different lexical and semantic variations of the ground-truth answer, and thus all of them are correct. We use only the test sets of these datasets for all our prompting experiments in a zero-shot setting. We use BioMRC LITE version of BioMRC. The statistics of the four datasets are listed in Table 1.

4 Prompting Techniques

While an exhaustive study of all prior prompting strategies could have been a better experimental setup but due to the cost-prohibitive nature of running large-scale experiments on GPT, we only select the techniques that have shown to perform well in the general domain. Along with these strategies, we also introduce a novel prompting method named *Implicit RAG*. We elaborate on all of these different prompts and their corresponding templates. There may be slight differences in prompt templates of the same prompting strategy across different datasets in order to adhere to the syntactic and semantic rules of English grammar as well as align with the dataset characteristics.


What is the name of the Health Provider?
Find sections in the document which may help in answering the question
Extract Relevant Sections
Answer From Extracted Sections

AUTHORIZATION FOR RELEASE OF HEALTH INFORMATION PURSUANT TO HIPAA

Patient Name	Date of Birth	Medical Record Number
Patient Address		

I, or my authorized representative, request that health information regarding my care and treatment as set forth on this form:

In accordance with New York State Law and the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), I understand that:

1. This authorization may include disclosure of information relating to **ALCOHOL** and **DRUG ABUSE, MENTAL HEALTH TREATMENT,** except psychotherapy notes, and **CONFIDENTIAL HIV* RELATED INFORMATION** only if I place my initials on the appropriate line in Item 9(a). In the event the health information described below includes any of these types of information, and I initial the line on the box in Item 9(a), I specifically authorize release of such information to the person(s) indicated in Item 8.
2. If I am authorizing the release of HIV-related, alcohol, or drug treatment, or mental health treatment information, the recipient is prohibited from redisclosing such information without my authorization unless permitted to do so under federal or state law. I understand that I have the right to request a list of people who may receive or use my HIV-related information without authorization. If I experience discrimination because of the release or disclosure of HIV-related information, I may contact the New York State Division of Human Rights at (212) 480-2493 or the New York City Commission of Human Rights at (212) 306-7450. These agencies are responsible for protecting my rights.
3. I have the right to revoke this authorization at any time by writing to the health care provider listed below. I understand that I may revoke this authorization except to the extent that action has already been taken based on this authorization.
4. I understand that signing this authorization is voluntary. My treatment, payment, enrollment in a health plan, or eligibility for benefits will not be conditioned upon my authorization of this disclosure.
5. Information disclosed under this authorization might be redisclosed by the recipient (except as noted above in Item 2), and this redisclosure may no longer be protected by federal or state law.
6. **THIS AUTHORIZATION DOES NOT AUTHORIZE YOU TO DISCUSS MY HEALTH INFORMATION OR MEDICAL CARE WITH ANYONE OTHER THAN THE ATTORNEY OR GOVERNMENTAL AGENCY SPECIFIED IN ITEM 9 (b).**

7. Name and address of health provider or entity to release this information:

8. Name and address of person(s) or category of person to whom this information will be sent:

9(a). Specific information to be released:
□ Medical Record form (insert date) ____________________to (insert date)____________________
□ Entire Medical Record, including patient histories, office notes (except psychotherapy notes), test results, radiology studies, films, referrals, consults, billing records, insurance records, and records sent to you by other health care providers.
□ Other:________________________________ Include: (*Indicate by Initialing*)
____
________ **Alcohol/Drug Treatment**
________ **Mental Health Information**
________ **HIV-Related Information**
________ **Genetic Testing**

Authorization to Discuss Health Information
(b). □ By initialing here _____ I authorize ____
Initials Name of individual health care provider
to discuss my health information with my attorney, or a governmental agency, listed here:
____
(Attorney/Firm or Governmental Agency Name)

10. Reason for release of information: □ At request of individual □ Other:	11. Date or event on which this authorization will expire:
12. If not the patient, name of person signing form:	13. Authority to sign on behalf of patient:

All Items on this form have been completed and my questions about this form have been answered. In addition, I have been provided a copy of the form.

______________________________ Date: ______________________________
Signature of Patient or representative authorized by law.

*** Human Immunodeficiency Virus that causes AIDS. The New York State Public Health Law protects information which reasonably could identify someone as having HIV symptoms or infection and information regarding a person's contacts.**

2. If I am authorizing the release of HIV-related, alcohol, or drug treatment, or mental health treatment information, the recipient is prohibited from redisclosing such information without my authorization unless permitted to do so under federal or state law. I understand that I have the right to request a list of people who may receive or use my HIV-related information without authorization. If I experience discrimination because of the release or disclosure of HIV-related information, I may contact the New York State Division of Human Rights at (212) 480-2493 or the New York City Commission of Human Rights at (212) 306-7450. These agencies are responsible for protecting my rights.

4. I understand that signing this authorization is voluntary. My treatment, payment, enrollment in a health plan, or eligibility for benefits will not be conditioned upon my authorization of this disclosure.

Figure 2: Implicit RAG Technique

Basic The prompt template used for this technique is shown in Figure 1. The Basic prompting approach asks GPT to answer the query in the simplest way possible. The *profession* placeholder specifies the role that GPT has to take in order to answer the asked question. Based on the source of the dataset, this placeholder takes the value of *biologist* in the case of ProcessBank, *biomedical researcher* in the case of BioMRC, *consumer healthcare expert* in case of MASH-QA and *medical expert* in the case of CliCR dataset. Again, based on the source of the dataset, the placeholder *context_type* takes the value of *paragraph* for ProcessBank, *abstract of the paper* for BioMRC, *healthcare article* for MASH-QA and *clinical case report* for CliCR. The placeholder *query_type* takes the value of *query* for ProcessBank, *title containing the missing entity* for BioMRC, *query* for MASH-QA and *query containing the missing entity* for CliCR. The *query_text* placeholder contains the actual text of the query and similarly *context_text* contains the actual text of the context. The *options* placeholder is present only for ProcessBank and BioMRC datasets and contains the choices to select from while answering the asked query.

Chain-of-Thought Reasoning (CoT) The rationale behind using the CoT technique is that there may be multiple smaller questions that need to be answered first in order to conclude the answer of the final asked question. For example, one of the questions asked to GPT is *Has there been at least 6 weeks of provider-directed conservative treatment?*. This question can easily be divided into 3 smaller questions *Has there been any conservative treatment?*, *Was the treatment provider-directed?* and *What was the duration of conservative treatment?*. The prompt template used for this technique is exactly same to that of Figure 1 with an additional line instructing the model to *Think step by step*.

Analogical Reasoning (AR) Inspired by Yasunaga et al. (2023b), we design our own analogical reasoning strategy by tweaking the prompt to fit our problem statement. We do this because, unlike the general domain, GPT would not be able to recall specific dataset-level knowledge, as we are not sure if it was ever trained on the datasets being used in our study. Rather, we frame the prompt so that GPT does not need to rely on a lot on global knowledge. To that end, instead of asking GPT to generate any kind of relevant QA pairs based on its global knowledge, we ask GPT to generate QA pairs from the given context and then answer the initial question. There is one hyperparameter for this technique which is the number of QA pairs to generate.

Implicit Retrieval Augmented Generation (RAG) Most of the work on RAG talks about data retrieval based on an accepted relevancy score and then using LLM prompts to answer the given query. The data retrieval is done by storing the embeddings from some encoder of the entire corpus (a datapoint's context in our case) in a vector database index and then retrieving the most match-

> You are a *{profession}* who is given a *{context_type}* and a corresponding *{query_type}*. Your job is to read the given *{context_type}* and then select the best option from a list of options to answer the *{query_type}*.
>
> The *{query_type}* that needs to be answered is listed below.
>
> *{query_type}*: *{query_text}*
> List of options: *options*
>
> Identify *{number_of_sections}* most relevant sections or text extracts from the given *{context_type}* that may help in selecting the best option to answer the given *{query_type}*. The identified sections or text extracts should be distinct from each other. The identified sections or text extracts must be between *{lower_limit_length}* to *{upper_limit_length}* words long.
>
> Now, choose the best option to answer the given *{query_type}* using the identified sections or text extracts.
>
> Here is the *{context_type}* that needs to be read to select the best option from a given list of options for the *{query_type}*.
>
> ### *{context_text}* ###

Figure 3: Implicit RAG Prompt Template

ing data points (text extracts or sections from a datapoint's context) for a given query. The key idea behind using RAG is that it helps in saving a lot of computational cost and improves the LLM's performance as now it has to look in a smaller knowledge space to answer the asked question. In our proposed novel prompting technique *Implicit RAG*, we completely ignore the overhead involved in getting the embeddings of the entire corpus and storing them in a vector database. Instead, we ask the LLM itself to find the most relevant text extracts or sections in the given context which may help in answering the asked question, and then later use these extracted sections to conclude the answer to the original question. The general working of our proposed prompting technique is shown in Figure 2. There are two hyper-parameters for this technique. First is, the number of sections to extract, and the next one is the number of words in each section or text extract. The prompt template used for this technique is shown in Figure 3. The hyper-parameter values for the number of sections and number of words in each section is provided in the placeholders *number_of_sections* and *lower_limit_length* and *upper_limit_length*.

5 Results & Analysis

[1]We use the 32k context window version of GPT-4 to conduct all our experiments. We set the temperature, frequency penalty, and presence penalty to 0 and max tokens to 1000 for GPT-4. The results for all the datasets have been discussed individually below. Based on different iterations of experiments, we choose the hyper-parameter number of QA pairs to generate for AR to be 3 for all the datasets. Similarly, for *Implicit RAG*, we choose the hyper-parameter *lower_limit_length* and *upper_limit_length* values as 50 and 200 respectively for all the datasets except MASH-QA. For MASH-QA, we choose *lower_limit_length* as 0 and *upper_limit_length* as 300. We choose *number_of_sections* for *Implicit RAG* to be 1 for MASH-QA, 2 for ProcessBank and 3 for BioMRC and CliCR.

ProcessBank The results for ProcessBank are shown in Table 2. We ran all 4 prompting strategies on the entire test set of 150 datapoints in a zero-shot setup. Every single prompting method outperforms the previously proposed methods, thus giving us

[1] We will release the relevant sections identified by the Implicit RAG technique as well as the question-answer pairs generated by the AR method upon acceptance for all the discussed datasets which can prove useful for other researchers.

Method	Accuracy
Basic (Full)	0.96
CoT (Full)	0.96
AR (Full)	0.96
Implicit RAG (Full)	**0.97**
Implicit RAG (Full)	**0.97**
Gold Structure	**0.77**
ProRead	0.67
SyntProx	0.60
TextProx	0.55
Bow	0.47

Table 2: Results on ProcessBank. The results for Gold Structure, ProRead, SyntProx, TextProx and Bow have been discussed in Berant et al. (2014b)

Method	Accuracy
Basic (1000)	**0.87**
CoT (1000)	0.81
AR (1000)	0.82
Implicit RAG (1000)	0.83
Basic (Full)	**0.87**
MLP-based Weighting	**0.88**
AoA-Reader with BioBERT	0.87
SciBERT-Max-Reader	0.80
AoA-Reader	0.70
AS-Reader	0.62

Table 3: Results on BioMRC. The results for models MLP-based Weighting and AoA-Reader with BioBERT are discussed in Lu et al. (2022) whereas the results for SciBERT-Max-Reader, AoA-Reader and AS-Reader are explained in Pappas et al. (2020)

a new SoTA on this dataset. Among the different prompting strategies, *Implicit RAG* gets the best results. The important observations are:

1. The only 4 to 5 datapoints that GPT got wrong either are very confusing for even humans to answer or had some typo or extra punctuation in ground-truth answers which GPT was not able to mimic during its generation.

2. It is observed that all the GPT prompting strategies work more or less the same if the question can be answered from a small span in the provided context. The reason why *Implicit RAG* is able to outperform other techniques is because this dataset includes around 30% of temporal and true-false type questions which require extensive analysis of the entire context and that the answer can be spread in different segments of the context. Therefore, reducing the knowledge space by extracting relevant sections to answer the asked question helps in improving the performance.

BioMRC The results for BioMRC are listed in Table 3. Due to cost-related reasons, we first compare different prompting methods by running them on a randomly selected 15% (1000 datapoints) subset of the test set and then choosing the best prompting technique to run on the entire test set. All these experiments are done in a zero-shot setting. Amongst the different prompting techniques, Basic prompting gets the best results and *Implicit RAG* ranks second. The important observations are:

1. Even though BioMRC is a cleaner version of BioREAD, there are still elements of lack of structure in the dataset. For example, there is no 1-1 mapping between entity IDs and entities. So, this means the same entity can be mapped to multiple entity IDs and vice versa which causes a lot of confusion when quantifying performance. The authors of BioMRC claim that for any query, the abstract or the context contains all the candidate options including the correct answer but this is not always true leading to more confusion during evaluation.

2. Quite a few times GPT is able to generate an acronym answer instead of its corresponding full form. Ideally, both acronyms and their full forms must be considered as correct answers.

3. There are instances where GPT being a generative model is able to produce semantically similar answers but still they are marked wrong as they do not exactly match with the correct answer. An embedding based metric can be helpful here.

4. There are a lot of entities which are semantically and syntactically the same but still belong to different ontologies and thus have different entity IDs. For example, in one case, GPT generates the answer *amino acids* where the correct answer is *amino acid* but since both of these entities have different IDs, this answer had to be marked wrong.

Another important aspect to note here apart from the lack of structure in the dataset is the overall system design of supervised models which are being compared to GPT when talking about SoTA. Supervised models use 70%-80% of the available data as their training set which allows their parameters to get a good idea about the nuances of the dataset whereas in case of GPT, all our experiments are being conducted in a zero-shot setting. Also since GPT is a generative model, the chances of GPT

Method	EM	F1	P	R
Basic (600)	**0.12**	**0.53**	0.50	**0.56**
Analogical (600)	0.11	0.50	**0.53**	0.47
CoT (600)	0.11	0.52	0.50	0.55
Implicit RAG (600)	0.10	0.52	0.51	0.52
Basic (Full)	**0.14**	**0.53**	**0.50**	**0.57**
Bert	0.9	0.25	0.56	0.16
RoBERTa	0.9	0.29	0.58	0.19
XLNet	0.9	0.29	0.56	0.20
MultiCo	**0.22**	**0.57**	**0.58**	**0.56**
Tanda	0.9	0.25	0.56	0.16

Table 4: Results on MASH-QA dataset. The results for Bert, RoBERTa, XLNet, MultiCo and Tanda have been talked about in Zhu et al. (2020)

Method	EM	F1
Basic (1100)	0.37	0.53
Analogical (1100)	**0.39**	**0.54**
CoT (1100)	0.36	0.51
Implicit RAG (1100)	0.38	**0.54**
Analogical (Full)	**0.34**	**0.52**
Human Novice	0.31	0.45
Human Expert	**0.35**	**0.54**
GA-Anonym	0.25	0.33
GA-Ent	0.22	0.30
GA-NoEnt	0.15	0.34
SA-Anonym	0.20	0.27
Sim-Entity	0.21	0.29

Table 5: Results on CliCR dataset. The results for Human Novice, Human Expert, GA-Anonym, GA-Ent, GA-NoEnt, SA-Anonym and Sim-Entity are explained in Šuster and Daelemans (2018)

generating an answer not present in the candidate answer list despite the final answer being semantically and syntactically the same is really high. But a supervised model is never going to face this problem as it makes its prediction based on the confidence score for each candidate answer and thus the final answer is always going to be present in the candidate answer list.

MASH-QA The results for MASH-QA are shown in Table 4. We start by conducting a comparison between different prompting strategies by evaluating them on a randomly selected 15% (600 datapoints) subset of the test set. These experiments are undertaken in a zero-shot setting. As we can see in Table 4, Basic prompting performs the best while *Implicit RAG* ranks second. The important points to discuss here are:

1. The answers in QA pairs of MASH-QA are very subjective. The authors of this dataset have not specified any structured process that was followed by the healthcare experts when trying to answer the questions asked on a website from where this dataset was sourced in the first place. An in-depth analysis shows that even though GPT is able to extract better answers a lot of times, since it does not match with the ground truth answers, the evaluation metrics do not reflect it's true capabilities.

2. There is a correlation observed between increase in the number of sections and decreasing performance of *Implicit RAG*. The reason is that the answers in this dataset are long span and thus with increasing number of sections, there is a loss of contextual continuity as the ground truth answers can get split across multiple sections. This ends up confusing the LLM making it difficult to choose the right set of sentences from different sections. Hence, it performs the best when asked to extract just one section from the context. But because MASH-QA contains answers which can be present in disjoint spans of the context, extracting just one section is not able to make *Implicit RAG* the best performing prompting method.

3. One question which may arise is whether extracting one section or having the hyper-parameter number of sections set to 1 for *Implicit RAG* makes it same as that of Basic prompting. *Implicit RAG* and Basic prompting become the same only when we are not only extracting just one section from the context but also when the hyper-parameter number of words for *Implicit RAG* is set equal to the length of the entire given context. But for MASH-QA, the hyper-parameter number of words is set to 300 with the lower limit being 0 and upper limit being 300 and hence they are different.

Again, we need to reiterate that GPT's performance in case of MASH-QA is being compared to supervised models which use 70%-80% of the total data as their training set allowing their parameters to capture granular details better than a generic LLM like GPT in a zero-shot setting.

CliCR The results for CliCR can be seen in Table 5. Again, due to cost-related reasons, we first compare different prompting techniques by running them on randomly selected 15% (1100 datapoints) subset of the test set and then choosing the best

Dataset	ProcessBank (50)		BioMRC (50)		MASH-QA (50)		CliCR (50)	
Pattern	✓(46)	✗(4)	✓(41)	✗(9)	✓(7)	✗(43)	✓(31)	✗(19)
Right Section	100%	100%	95%	56%	100%	93%	81%	32%
Wrong Section	0%	0%	5%	44%	0%	7%	19%	68%

Table 6: Qualitative Analysis of *Implicit RAG* on ProcessBank, BioMRC, MASH-QA and CliCR

prompting method to run on the entire test set. All these experiments are done in a zero-shot setting. Amongst the different prompting strategies, *Implicit RAG* and AR get the best results in terms of F1 metric. AR minutely performs better than *Implicit RAG* when compared in terms of the Exact Match (EM) metric. However, EM is a very harsh metric for a generative model as there can be so many possible variations of semantically similar output which are not wrong. Since AR is computationally faster with respect to *Implicit RAG*, we ran AR on the entire dataset. All the prompting methods outperform the previously proposed methods. Not only does GPT surpass the performance of previous models, but it also comes close to Human Expert performance while beating Human Novice results. The important observations are:

1. The authors of CliCR mention that for the training of supervised models, only those instances are used for which at least one ground-truth answer from the set of ground-truth answers occurs in the clinical case report or the context. But for the evaluation part, for both validation and test sets even those datapoints are included where there is no intersection between ground-truth set and the entities mentioned in the context. This favors supervised learning settings as supervised models have a separate training and development set which can allow the parameters of the model to learn such cues. GPT is still able to perform better possibly because the global knowledge embedded in its parameters gives it enough evidence to perform well.

2. The authors of CliCR compare various skills in their work between the previous SoTA (GPT is the new SoTA) model GA-NoEnt and Human Expert and show that there still exists a huge gap between them. Since GPT is able to achieve almost Human Expert level performance, we can expect it to show similar capabilities in other MRC tasks.

3. There are multiple reasons why *Implicit RAG* performs well on this dataset. First, the mean length of context in this dataset is 1461 words which indicates that with increasing size of context, the chances of analysis of different sections of the context simultaneously to answer a question is high and that is the core idea behind *Implicit RAG*. Next, the authors of CliCR list out that 70% of the queries in this dataset require the *bridging* skill, 40% require the skill of *tracking* and around 25% demand the *spatiotemporal* skill. All these three skills indicate that answering queries in this dataset require deriving cues from different segments of the context and that is what we propose as the key rationale behind *Implicit RAG*.

Implicit RAG Out of the four datasets that we use in this study, *Implicit RAG* is able to achieve the best results for two of them when compared with other prompting techniques. It ranks at the second place for the other two datasets. One of the questions which may arise is whether Implicit RAG can be applicable to contexts which cannot fit in LLM's 32k token limit. Implicit RAG will especially perform better than other prompting techniques in cases where the context size is more than 32k. In such cases, we can chunk the context and make multiple calls to Implicit RAG to retrieve relevant sections given the query. Once all the relevant sections have been retrieved, the last call to Implicit RAG can use these sections to arrive to an answer. But all other prompting techniques require analysis of the entire context (greater than 32k in this case) at the same time to arrive to an answer. We further do a qualitative analysis on 50 randomly picked datapoints for all four datasets. We check how many times the extracted sections are relevant to the question or not. Even if 1 out of all the extracted sections are relevant, we consider that to be a valid retrieval irrespective of whether the final answer was correct or incorrect. The results are shown in Table 6. As we can see, *Implicit RAG* is able to extract relevant sections in most cases.

6 Conclusion

In this work, we show that even in a zero-shot setting, GPT surpasses the performance of supervised models for two out of four benchmarks. Furthermore, GPT's performance comes close to that of Human Expert for one of the benchmarks. Our study corroborates that LLMs indeed have surpassed preconceived techniques even on difficult to model domains like biomedicine. We also come up with a novel prompting method *Implicit RAG* which gets the best results in two out of four datasets and ends up at rank two in others. This opens a new research direction for the RAG domain allowing other researchers to experiment with this technique on other domain datasets.

7 Limitations

Due to cost associated with running large-scale experiments with GPT, we did a comparison of different prompting techniques on a subset of about 15% of the entire test set for three out of four datasets we discuss in this work. It could be possible that there may be a slight difference in the distribution of the random subset we chose in comparison to the entire test set and this could potentially change the final results obtained by a given prompting technique although we expect the difference to be small. As discussed earlier, in cases where the answer to a query can be found in a small span of the context, there is not a huge difference between different prompting techniques. Thus, running the Basic prompting method will be computationally more inexpensive than running heavier prompting strategies like AR or *Implicit RAG*.

References

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014a. Modeling biological processes for reading comprehension. In *Conference on Empirical Methods in Natural Language Processing*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014b. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1499–1510.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Kunming Cheng, Qiang Guo, Yongbin He, Yanqiu Lu, Ruijie Xie, Cheng Li, and Haiyang Wu. 2023. Artificial intelligence in sports medicine: could gpt-4 make human doctors obsolete? *Annals of Biomedical Engineering*, pages 1–5.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuxuan Lu, Jingya Yan, Zhixuan Qi, Zhongzheng Ge, and Yongping Du. 2022. Contextual embedding and model weighting by fusing domain knowledge on biomedical question answering. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–4.

Maria Mahbub, Sudarshan Srinivasan, Edmon Begoli, and Gregory D Peterson. 2022. Bioadapt-mrc: adversarial learning-based domain adaptation improves biomedical machine reading comprehension task. *Bioinformatics*, 38(18):4369–4379.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Dimitris Pappas, Ion Androutsopoulos, and Harris Papageorgiou. 2018. Bioread: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. Biomrc: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel R Bowman. 2022. What makes reading comprehension questions difficult? *arXiv preprint arXiv:2203.06342*.

Simon Šuster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. 2023. Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. *arXiv preprint arXiv:2308.06294*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2023a. Large Language Models as Analogical Reasoners. ArXiv:2310.01714 [cs].

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023b. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849.