Natural Language Processing Journal 8 (2024) 100083

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset

Zafaryab Rasool [a,*], Stefanus Kurniawan [a], Sherwin Balugo [a], Scott Barnett [a], Rajesh Vasa [a], Courtney Chesser [b], Benjamin M. Hampstead [c,d], Sylvie Belleville [e,f], Kon Mouzakis [a], Alex Bahar-Fuchs [b]

[a] Applied Artificial Intelligence Institute, Deakin University, Melbourne, Australia
[b] School of Psychology, Faculty of Health, Deakin University, Melbourne, Australia
[c] Mental Health Service, VA Ann Arbor Healthcare System, Ann Arbor, USA
[d] Research Program on Cognition and Neuromodulation Based Interventions, Department of Psychiatry, University of Michigan, Ann Arbor, USA
[e] Psychology Department, Université de Montréal, Quebec, Canada
[f] Research Center, Institut Universitaire de Gériatrie de Montréal, Canada

ARTICLE INFO

Keywords:
Large language models
Document-based information retrieval
Question-answering
Retrieval augmented generation
CogTale dataset
Healthcare

ABSTRACT

Document-based Question-Answering (QA) tasks are crucial for precise information retrieval. While some existing work focus on evaluating large language model's (LLMs) performance on retrieving and answering questions from documents, assessing the LLMs performance on QA types that require exact answer selection from predefined options and numerical extraction is yet to be fully assessed. In this paper, we specifically focus on this underexplored context and conduct empirical analysis of LLMs (GPT-4 and GPT-3.5) on question types, including single-choice, yes–no, multiple-choice, and number extraction questions from documents. We use the CogTale dataset for evaluation, which provide human expert-tagged responses, offering a robust benchmark for precision and factual grounding. We found that LLMs, particularly GPT-4, can precisely answer many single-choice and yes–no questions given relevant context, demonstrating their efficacy in information retrieval tasks. However, their performance diminishes when confronted with multiple-choice and number extraction formats, lowering the overall performance of the models on this task, indicating that these models may not yet be sufficiently reliable for the task. This limits the applications of LLMs on applications demanding precise information extraction and inference from documents, such as meta-analysis tasks. Our work offers a framework for ongoing dataset evaluation, ensuring that LLM applications for information retrieval and document analysis continue to meet evolving standards.

1. Introduction

Large language models (LLMs) have recently gained attention due to their ability to solve various natural language processing tasks (Espejel et al., 2023; Aher et al., 2023; Acharya et al., 2023; Rasool et al., 2024; Zhao et al., 2023). However, existing evaluation of LLMs predominantly focuses on general knowledge questions and reasoning tasks (Bian et al., 2023; Qin et al., 2023; Bai et al., 2023; Bang et al., 2023), rather than retrieval of specific information from documents. In real-world, various scenarios require extracting the number of participants in the control group of a trial in a medical paper, relevant policy information from policy documents, specific dollar value of liability in a legal context, and so on. The current favored approach to solve these tasks involve using Retrieval Augmented Generation (RAG, Lewis et al., 2020). However, the effectiveness of LLMs on these narrow tasks is under explored, which limits the evaluation of these system's real-world applicability. Furthermore, while existing datasets tend to focus on the performance of LLM, document QA datasets around RAG are in their infancy.

We take the example of the CogTale platform (Sabates et al., 2021) as a running example which consist of database of published research papers on cognitive interventions for older adults. Researchers interested in evaluating the quality of trials entered onto the CogTale database and synthesizing the evidence generally perform manual annotation and data extraction into the database using a structured form. However, the manual retrieval/extraction of target information from these documents is a laborious process, potentially leading to

* Corresponding author.
E-mail addresses: zafaryab.rasool@deakin.edu.au (Z. Rasool), stefanus.kurniawan@deakin.edu.au (S. Kurniawan), s.balugo@deakin.edu.au (S. Balugo), scott.barnett@deakin.edu.au (S. Barnett), rajesh.vasa@deakin.edu.au (R. Vasa), courtney.chesser@deakin.edu.au (C. Chesser), bhampste@med.umich.edu (B.M. Hampstead), sylvie.belleville@umontreal.ca (S. Belleville), kon.mouzakis@deakin.edu.au (K. Mouzakis), a.baharfuchs@deakin.edu.au (A. Bahar-Fuchs).

https://doi.org/10.1016/j.nlp.2024.100083
Received 30 December 2023; Received in revised form 15 May 2024; Accepted 6 June 2024

Table 1
Example questions from the CogTale dataset belonging to different question-type category.

Question: *Was the intervention delivered as per the planned protocol? i.e., no significant changes to the protocol implemented after the trial began?* Category: Yes-No type Options: [Yes, No, Not specified] Actual answer:
Question: *What type of trial was conducted to evaluate the intervention?* Category: Single-choice Options: [Randomised controlled trial- parallel groups, Randomised controlled trial- cross over trial, Randomised controlled trial- cluster, Randomised controlled trial -Waitlist-control, Non randomised controlled trial, Open (before and after) trial (no control), Single case (with phase randomization), Single case (without phase randomization), Randomized interventional study (no control group), Partial-randomized controlled trial, Parallel groups] Actual answer:
Question: *What is the number of control conditions?* Category: Single-choice (number) Option: [0, 1, 2, . . . , 21] Actual answer:
Question: *Which individuals were deliberately kept unaware of the specific intervention they received in the study?* Category: Multiple-choice Option: [Assessors, Trainers/therapists, Participants, Data analysts, No blinding attempted, Not specified, N/A, Caregivers] Actual answer:
Question: *What proportion of participants from the control group were retained at the post-intervention assessment?* Category: Number-extraction Actual answer:

challenges such as mis-interpretation, scalability issues for projects with stringent timelines, inconsistencies, and the potential for errors, which slows down the evidence translation and implementation process. LLMs which have proved their effectiveness on several tasks such as summarizing, reasoning and others, can offer potential solution to the aforementioned issues. Therefore, there is a need to investigate the performance of LLMs in information retrieval tasks.

Existing related work by Pereira et al. (2023) evaluated GPT-3's performance on the above task using three datasets (IIRC, Qasper and StrategyQA). These dataset mostly focus on complex context comprehension and multi-paragraph answer extraction. Other popular datasets such as PubMedQA (Jin et al., 2019) and BioASQ (Krithara et al., 2023) involve either asking yes–no type question or factoid and list questions. However, how LLMs perform on question types that require selecting answers from provided response options and extracting numerical values is not yet fully explored. Such question–answer formats are prevalent in various scenarios, including in healthcare-related evidence synthesis tasks, and gaining insights into LLMs performance in these areas would enable users to confidently employ them for such tasks. These questions may also require inferring answers from the context, and answers may not be directly stated. Examples of such questions are shown in Table 1.

Therefore, in this paper, we focus on the task of extracting and inferring specific information from the CogTale dataset using LLM, specifically GPT-3.5-turbo and GPT-4 (OpenAI, 2023). We developed a pipeline which involves extracting related passages from document(s) based on the question, and prompting an LLM to select the correct answer(s) from a set of options using the extracted passages. CogTale data extraction form consists of a set of questions that can be categorized into single choice, multiple choice, single choice (number options), yes–no type, etc. Additionally, direct value/number extraction and value computation questions are also included. These questions along with related passages extracted from the document(s) are passed to an LLM for generating the answers.

We conduct an empirical analysis on 13 studies, consisting of the research papers and the different question types selected from the Cogtale platform, using the above developed pipeline and various prompting techniques. Based on the analysis, we found that GPT-4 surpassed GPT-3.5-turbo in performance across all question types from CogTale dataset. However, the overall performance of these models was not found satisfactory as GPT-4 achieved an overall accuracy of 41.84%. In terms of the different categories of questions, GPT-4 performed better on single-choice questions and yes–no type questions as compared to multiple-choice and number-extraction. Our further exploration with different prompting techniques resulted in slight improvement on yes–no and single-choice categories. Additionally, we investigated the retriever's performance on various incorrect responses and observed that the models struggled to infer or select correct options, even when the relevant information was present in the extracted chunks/passages. This shows the challenges these models face in tasks requiring inference from context and accurate answer selection, demonstrating that the current versions of GPT may not be sufficiently reliable for the task. Our study underscores the necessity for more robust strategies and evaluation methodologies to overcome the identified limitations. By doing so, we can enhance the reliability and applicability of language models for a wide range of QA tasks.

We summarize our contributions in this paper as follows:

- **Diverse Question formats:** We conducted experimental analysis of Large Language Models (LLMs) with a focus on GPT-4 and GPT-3.5-turbo across diverse question formats such as yes–no, single-choice, multiple-choice, number-extractions using various prompting techniques, in the context of document-based information retrieval.
- **Utilization of CogTale dataset:** Leveraged the CogTale dataset, featuring research papers on cognitive interventions for older adults, to demonstrate the practical applicability of LLMs in retrieving information from documents, offering valuable insights for researchers and practitioners in healthcare and related fields.

In the remainder of the paper, we first discuss the background in Section 2 and then present the Methodology in Section 3 covering the details of the dataset and the QA framework. Empirical evaluation and analysis are discussed in Section 4. We provide a discussion of the results and future work in Section 5. Finally, we present the conclusion in Section 6, and threat to validity in Section 7.

2. Background

A notable surge in research endeavors has been directed towards the exploration of large language models recently. This surge has seen numerous studies evaluating LLMs performance on different tasks as highlighted in recent surveys (Zhao et al. (2023), Chang et al. (2023), Kalyan (2023)). Most existing works evaluate the performance

Table 2
Comparison of different QA datasets based on their question-types. Here '–' indicates that the particular category type is not present or not covered specifically in the dataset.

Dataset	Single-choice	Multiple-choice	Single-choice (numbers)	Yes-No	Number extraction
IIRC	–	–	–	–	–
StrategyQA	–	–	–	Yes	–
Qasper	–	–	–	Yes	–
PubMedQA	–	–	–	Yes	–
BioASQ	–	–	–	Yes	–
CogTale	Yes	Yes	Yes	Yes	Yes

of LLMs on benchmark and open-domain questions focused on reasoning and factoid questions. Bian et al. (2023) evaluated the performance of ChatGPT on commonsense problems from different domains. Qin et al. (2023) investigated the zero-shot performance of ChatGPT and GPT 3.5 on several NLP tasks. Bai et al. (2023) propose to use the language model as a knowledgeable examiner which evaluates other models on the responses to its questions. Bang et al. (2023) evaluates ChatGPT on NLP tasks. Kamalloo et al. (2023) evaluates LLMs and other open-domain QA models by manually evaluating their answers on a benchmark dataset.

Information retrieval using LLMs have gained attention recently (Ram et al., 2023; Shi et al., 2023; Levine et al., 2022). A recent work by Pereira et al. (2023) evaluates GPT-3's performance on three information-seeking datasets including the Incomplete Information Reading Comprehension (IIRC) Questions dataset (Ferguson et al., 2020), QASPER dataset (Dasigi et al., 2021) and StrategyQA dataset (Geva et al., 2021). Liu et al. (2023) also evaluated their approach on StrategyQA dataset. While the IIRC and StrategyQA dataset are focused on complex context comprehension and multi-paragraph evidence extraction, QASPER dataset consist of research papers focused on natural language processing topics and question types such as: Extractive, Abstractive, Yes/No and Unanswerable.

CogTale dataset differs from the above datasets and other popular datasets such as PubMedQA and BioASQ. PubMedQA (Jin et al., 2019) focuses on biomedical research papers and uses the abstract of their research papers for questions, while CogTale uses data extracted/retrieved from complete papers. BioASQ (Paliouras and Krithara, 2014) focus on open-domain QA over PubMed abstracts and include yes–no type, factoid and list questions. Cogtale specifically focuses on question-types such as selecting single or multiple correct answer(s) from a list of options specially, which differentiates it from other existing datasets. Table 2 provides a summary of the comparison of the above datasets including the Cogtale dataset. While the datasets may involve some questions of these types, they are not specifically focused on such types. Thus, bridging this gap by evaluating LLMs performance on CogTale dataset is crucial as it addresses a fundamental aspect of information retrieval, enabling context-aware and efficient access to knowledge from documents.

3. Methodology

We first describe the details of the CogTale dataset and then discuss the framework to evaluate the performance of LLMs on question-answering tasks.

3.1. Details of CogTale dataset

The CogTale platform is a repository of methodological and outcome data from trials of cognition-oriented treatments for the elderly and was developed as part of efforts to semi-automate key aspects of the evidence synthesis pipeline. CogTale serves as a valuable resource for researchers and clinicians seeking related information about trials and/or interested in rapid evidence synthesis. Facilitated by the CogTale platform, users can seamlessly search for specific studies and access precise details from them. Furthermore, the platform enables users to contribute their own studies, establishing an efficient medium for information retrieval. The platform include a wealth of data for each study included in the dataset, such as trial specifications, total sample size and its rationale, eligibility criteria, primary and secondary outcomes, intervention particulars, study findings, and more. Next, we discuss the question-types.

Question types: The CogTale data extraction form comprises a diverse array of questions, organized into eight different sections, focused on very specific information from different studies (i.e., research papers). The same question set is used to extract information from all studies in the database and seek information about how a trial was conducted, number of participants, and other useful information. We classify these questions into distinct types, providing detailed elaboration below.

1. **Yes-No type**: This question category requires a response in the form of 'yes" or 'no". This category may also include options such as 'Not Specified", 'N/A" or other similar options.
2. **Single-choice**: The second type involves questions accompanied by multiple options, with a singular correct answer among them.
3. **Single-choice (number)**: This category of questions is a subtype of single-choice questions, where the options consist solely of numerical values, with only one option being correct.
4. **Multiple-choice**: In this category, questions present multiple options, with the possibility of more than one option being correct.
5. **Number-extraction**: In this category, the expected response is a numerical value. However, this category does not involve options which distinguishes it from the third category: single-choice (number).

Examples of these question type from the CogTale dataset are shown in Table 1.

3.2. Question-answering (QA) framework

We explain the QA framework using document-based QA dataset (particularly CogTale dataset). The task of retrieving specific information from CogTale dataset can be described as the below problem definition. *Given a question q with a list of options and a document d, use the LLM to select the answer to question q utilizing information from d as supporting context.*

Based on the above problem definition, the QA framework is illustrated as shown in Fig. 1. Broadly there are two steps in this framework: (1) Retrieve, and (2) Answer. In the first step, a retriever is responsible for retrieving the most relevant information from the document based on the input question. This is an important step as it ensures that the model considers the appropriate context when generating answers. Initially, the document is divided into chunks, and then embeddings are generated for each chunk using an embedding model. When a new question comes, the retriever uses the question embeddings to find the relevant document chunks. This is a similarity task as most similar chunks to the question are required. Thus, an appropriate similarity measure (such as Cosine Similarity) is used to extract the most relevant chunks.

The selected chunks, along with the question, are passed to an LLM (GPT-3.5-turbo and GPT-4) to generate the answer. Since most questions in our study focus on selecting answer(s) from options, the

Table 3
Prompt example: Single-choice.

Use the following pieces of context to extract the information at the end. If you can't find the answer, just say that you don't know, don't try to make up an answer. {summaries} You can only answer from one of these values: {answer_options} Question: {question} The answer can only be the exact value of one of the options. Just return the final value when answering. Answer:

Table 4
Prompt example: Multiple-choice.

Use the following pieces of context to extract the information at the end. If you can't find the answer, just say that you don't know, don't try to make up an answer. {summaries} You can only answer from any of these values: {answer_options} Question: {question} You can pick many options from the provided options, but you can only use each once. Return the final values when answering. Answer:

Table 5
Prompt example: Number-extraction.

Use the following pieces of context to extract the information at the end. If you can't find the answer, just say that you don't know, don't try to make up an answer. {summaries} Question: {question} The answer can only be a number (or decimals) value between 0 to 1. Just return the final value when answering. Answer:

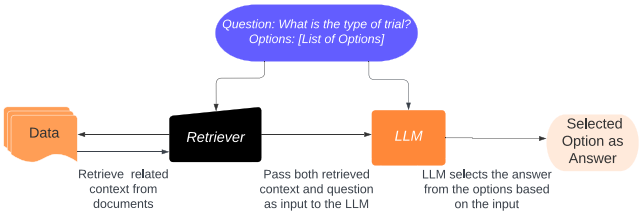

Fig. 1. Question-Answering (QA) Framework using LLM for a document-based QA task.

LLM is required to select one or multiple correct answer based on the question-type. For this task, an LLM needs to be prompted and appropriate prompting is essential. For this study, we utilize straightforward prompts that explicitly outline the question's requirements. Our prompts to the model consist of a question presented alongside answer options, as well as the corresponding passages extracted from the document. For the different question types we discussed earlier, the example of prompts are shown in the Tables 3–5. Additionally, we also use other prompting strategies for evaluation.

4. Evaluation and results

In this section, we describe the framework and the evaluation details, along with the results and analysis.

4.1. Evaluation details

We performed empirical analysis on 13 studies covering 337 questions and response pairs from the CogTale dataset, and compare the generated answers using the LLMs with the actual responses.

Framework Details: For testing our QA framework on the CogTale dataset, we use the Retrieval Augmented Generation (RAG) pipeline described by Barnett et al. (2024). In this RAG pipeline, we first generate the embeddings from text chunks extracted from the documents in our dataset using the OPENAI model (text-embedding-ada-002). Then, we employ the FAISS library (Johnson et al., 2019) which allows efficient indexing of these embeddings and similarity search of vectors. Given a query, we generate its embedding and retrieve relevant chunks to the query from the FAISS index using the FAISS *similarity_search_with_score* method. This method retrieves the top relevant or most similar chunk embeddings for the given query. Finally, using the query and the retrieved chunks as input, an LLM is prompted to generate the answer. These retrieved chunks serve as context for the LLM to answer the query.

Study Selection: The CogTale dataset comprises studies categorized as verified and unverified, with verified studies encompassing published research papers. Among these, 52 studies were identified as verified. During manual scrutiny, we excluded studies where the correct answer was not provided among the options. Subsequently, to enhance the internal validity of our analysis and streamline the complexity, studies with more than one intervention or control component were excluded. Following these filtering steps, 40 studies remained. These were further divided into three sets, each containing 13, 13, and 14 studies. This subdivision aimed to analyze the sets separately and comprehend the model's performance on each. For this particular study, we focused on one of the set comprising 13 studies, reserving the others for future investigations. The titles of these selected studies are given in Table 9 in Appendix.

Among these 13 studies, most of the studies comprise 1 document, except one which consist of two documents. For each of the study, the dataset consist of 28 questions of various types requiring specific information. However, for a few studies that we used, some of the questions were not pertinent to the study and we do not use them for evaluation. For instance, if a study is not of the randomized control type, specific questions lack a ground truth, and we omit asking these questions. Consequently, for each study, we pose only those questions for which the answer is present in the study. This approach resulted in a total of 337 question and response pairs evaluated across all the studies.

Table 6
Accuracy (in %) of GPT-3.5-turbo and GPT-4 on CogTale dataset on the different type of questions. Here, Questions (%) represents percentage of particular question-type. N represents the number of questions of the specific categories.

Category of questions	N	Questions (%)	GPT-3.5	GPT-4	GPT-3.5 (CoT)	GPT-4 (CoT)	GPT-3.5 (Few-Shot)	GPT-4 (Few-Shot)
Yes-No	143	42.43%	41.96%	46.85%	39.16%	49.65%	44.76%	**52.45**%
Single-choice	73	21.66%	38.36%	**56.16**%	43.84%	45.21%	31.51%	46.58%
Single-choice (number)	52	15.43%	17.31%	32.69%	17.31%	28.85%	38.46%	**59.62**%
Multiple-choice	43	12.76%	4.65%	**25.58**%	2.33%	16.28%	0.00%	0.00%
Number-extraction	26	7.72%	**26.92**%	19.23%	3.85%	0.00%	3.85%	11.54%
Overall Results	337	100%	31.45%	41.84%	29.38%	37.39%	32.05%	**42.43**%

Models: We selected GPT-3.5-turbo and GPT-4 models to evaluate the document-based QA tasks on different question-types. These models are known for their natural language processing capabilities. GPT-3.5-turbo is known for its cost-effectiveness and efficient use of resources compared to some larger models, making it a practical choice for certain applications. On the other hand, GPT-4 represents a more recent and potentially more sophisticated iteration, offering an opportunity to assess the advancements in performance. We set the temperature variable of the models to 1 to ensure consistency in the analysis and results interpretation.

Metric: We conduct a comparison between the answers generated by the models and the actual ground truth answers, reporting whether they align. Given that the answers fall into either multiple-choice options or single numerical values, we perform a direct comparison between the generated answer and the ground truth, determining accuracy based on matching results. It is important to note that the ground truth results for the questions are provided within the dataset, offering a reliable reference for evaluation.

4.2. Results and analysis

In this section, we report the performance of models on the different question types, impact of prompting techniques and the impact of retriever on the performance of models.

Performance of Models: We first analyze the performance of the LLMs using the straightforward prompting (discussed earlier) as shown in fourth and fifth column of Table 6. The results in the last four columns use advanced prompting techniques and are discussed later under Impact of Prompting Techniques in this Section.

The overall accuracy of GPT-3.5-turbo and GPT-4 on these questions was found to be 31.45% and 41.84%, which shows GPT-4 outperformed GPT-3.5-turbo on the different question types in terms of correctly answering the questions. However, the overall accuracy of the two models is very low. The performance of these models across the different question categories is shown in Table 6. The best results are shown in bold. As shown, GPT-4 consistently outperformed GPT-3.5-turbo across various categories. Notably, both models exhibited superior performance in answering Yes-No and Single-choice type questions compared to other question types. However, their performance was less satisfactory for the Multiple-choice and Number Extraction question types.

We look into some examples of yes–no type questions where the model failed to answer correctly. For example, in Table 7, in the first question of yes–no type, while the actual answer is 'Yes', the models select 'Yes-Partially described" as the answer. It is possible, the generated answer may have been correct when only yes and no options were present. However, in the second and third questions, it did not correctly answer the question.

For single-choice questions shown in the table, we found that GPT-3.5 selected options that were not present in the list of options. Similarly, GPT-4 also selected 'Visual Imagery' option as the correct answer (for the fourth question) that was not present in the option list. These results are due to models hallucinating answers. Additionally, when only one of the option is to be selected for the question *What was the primary target of the intervention?*, GPT-3.5 instead answers the question in detail. The examples of single-choice (numbers) and number-extraction questions are shown in the last 4 columns in Table 7. These results show that the models have difficulty in answering and inferring numerical questions.

On multiple-type, GPT 3.5-turbo achieved very low accuracy (4.65%) as compared to GPT 4 (25.58%), demonstrating better reasoning abilities of GPT-4 on such questions. However, both the models achieved very low accuracy score as compared to single-choice categories. This is mainly due to both GPT-3.5 and GPT-4 models selecting more options than the actual number of correct answers, which causes this category to perform poorly. Table 8 show examples of such cases. For this question, the options consist of multiple potential answers. Despite only one option being correct for the question, both the models identified several options as correct. Additionally, it is important to highlight that the options selected by these models exhibited variations both in number and order.

Impact of Prompting Techniques: We further investigated the impact of popular prompting techniques such as Zero-shot Chain of Thought (CoT — (Kojima et al., 2022)) and Few-Shot (Brown et al., 2020) for answering questions from the CogTale dataset. These prompting techniques have shown to improve model's response in various scenarios (Kojima et al., 2022; Lee et al., 2024; Singhal et al., 2023). Zero-shot CoT simply ask the model to think step by step before producing any response. It generates the rationales for each task and follows these rationales to generate the final answer. On the other hand, Few-Shot prompting appends examples/demonstrations to the prompts to indicate how the model should perform. We used 5 demonstrations for the analysis.

The results of employing these techniques with both models are presented in the last four columns of Table 6. Zero-Shot CoT resulted in a slight overall accuracy decrease, while Few-Shot prompting yielded slightly higher accuracy rates. Notably, Zero-Shot CoT exhibited minimal impact on model performance across various question categories, except number-extraction. Conversely, Few-Shot prompting showcased enhanced performance in Yes-No and Single-choice (number) questions but demonstrated considerably lower accuracy in multiple-choice and number-extraction question types. On examining questions that were initially answered correctly using the straightforward prompting, we found that for the multiple-choice questions, the models selected different or more options as answers which lead to low accuracy on this category. On the other hand, for number-extraction questions, the models either provided a different answer or responded 'I don't know', which suggest that the model could not infer the correct answer from the retrieved chunks.

It is worth noting that the improvements in some question categories may be attributed to subtle changes in prompts rather than solely to enhanced reasoning capabilities.

Impact of Retriever: It is also imperative to test the relevance of the extracted chunks by the retriever as the LLMs can only be expected to generate accurate responses provided relevant chunks are retrieved from the documents. For retrieving the top relevant chunks for a query i.e., the top-k chunks, we set the value of $k = 4$ (default value).

We performed a manual evaluation to analyze the retrieved chunks for each of the first 100 incorrect responses generated by GPT-4, to provide insights into the performance of the retriever.

Table 7
Example of different question-type where the model generated incorrect answer. The questions shown here are randomly selected and are from different studies.

Question	Category	Options	Actual answer	GPT 3.5	GPT-4
Were the source(s) of participant recruitment adequately described?	Yes-No type	['Yes', 'No', 'N/A', 'Yes- Fully described', 'Yes- Partially described']	Yes	Yes- Partially described	Yes- Partially described
Does it mention any additional training given to participants about different strategies to use to perform and learn in the task?	Yes-No type	[Yes- single strategy, Yes- multiple strategies, No, N/A, Unclear, Yes]	No	Yes- multiple strategies	Yes- multiple strategies
What is name of the intervention?	Single-choice	[Cognitive training, Cognitive stimulation, Cognitive rehabilitation, Brain training, Brain games, Cognitive exercise, Mind training, Memory training, Memory games, Memory rehabilitation, ..., Multi-Strategic Metamemory Training (MMT), Computerized Cognitive Training]	Memory training	PM training	Cognitive training
What was the primary target of the intervention?	Single-choice	[Standardised cognitive task performance, Standardised skill performance, Everyday activity in context (e.g., banking), Psychoeducation, ... Memory and learning, Stress, ..., Not specified, Speed/accuracy of processing]	Memory and learning	The primary target of the intervention was the prospective memory (PM) component, specifically the enhancement of the prospective cue detection.	Visual imagery
What is the proportion of participants retained in the post-treatment experimental group?	Number-extraction	–	0.88	0.88	0.76
What is the proportion of participants retained in the post-treatment experimental group?	Number-extraction	–	0.80	0.75	0.80
What is the number of treatment/experimental conditions (not including control conditions)?	Single-choice (numbers)	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 16, 18, 27, 19, 14, 20, 23, 24, 21]	1	I don't know.	1
What is the number of control conditions?	Single-choice (numbers)	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 16, 18, 27, 19, 14, 20, 23, 24, 21]	1	80	80

We found that for 62 out of 100 questions, their corresponding retrieved chunks contained the relevant information to answer the question, while for the 38 questions, the corresponding retrieved chunks were not relevant to the question. This indicates a significant gap in the language model's ability to effectively utilize the retrieved information for inferring and generating accurate answers. Despite having access to relevant context in 62% of the cases (questions), the model still failed to provide correct answers. This underscores potential challenges such as understanding context, reasoning, or generating accurate responses, which need to be addressed to improve the model's performance in question answering tasks.

For the remaining 38 questions, the retriever could not identify the top relevant chunks from the document. This highlights potential limitations in the retrieval process and require evaluation of other different approaches. Additionally, the model still provided an answer for most of these questions, which is another limitation with this method.

5. Discussion and future work

Our study on the CogTale dataset reveals that GPT-4 surpasses GPT-3.5-turbo in question-answering accuracy, achieving 41.84% overall accuracy compared to 31.45% using straightforward prompts. These low accuracy percentages indicate unreliable performance of the LLMs on this task, prompting further investigation into whether inference and answer selection from options can be improved using different prompting techniques. However, even with advanced prompting strategies such as Zero-shot CoT and Few-Shot, there was little improvement observed in the overall performance. We will now dig deeper into the performance of the models on different question categories and discuss the performance of the retriever used.

In answering various question categories with straightforward prompts, both models perform well in Yes-No and Single-choice questions but encounter difficulties with Multiple-choice and Number Extraction types. GPT-4, while an improvement, faces challenges in nuanced understanding and occasionally provides incorrect answers in Yes-No questions. For some single-choice questions, these models were observed to choose options that were not part of the provided set of options, resulting in inaccuracies, possibly due to hallucination. Multiple-choice questions pose difficulties, with both models extracting more answers than necessary. GPT-3.5-turbo's low accuracy in Multiple-choice questions highlights limitations. Numerical understanding poses a distinct hurdle, reflected in the difficulty these models encountered in accurately inferring and answering numerical type questions. Furthermore, results indicate that using Few-Shot prompting improved performance of models on single-choice and yes–no category of questions. These investigations focused on GPT-3.5 and GPT-4, and highlighted the issues that the future language models should address by emphasizing nuanced comprehension and context awareness. Future work could explore the performance of other language models such as

Table 8
Example of Multiple-choice from options question-type from a study, where both GPT 3.5 and GPT-4 selected more options as correct answer than the actual number of answer.

Question:	Options	Actual answer	GPT 3.5	GPT-4
What were the other targets of the intervention?	[Standardised cognitive task performance, Standardised skill performance, Everyday activity in context (e.g., banking), Psychoeducation, Goal-setting & performance, Pleasant activities, Reminiscence, Reality orientation, Relaxation, Conversation, N/A, well being, Mood, Physical Activity, Quality of Life (QoL), Subjective Cognition (e.g., memory), Activities of daily living, Stress management, Assertiveness, Use of memory aids, fMRI Imaging, Paper-and-pencil exercises, Behaviour, Global functioning, Social interaction, Self-confidence, Motivation, Psychiatric symptoms, Emotional symptoms, Strategy, Global cognitive function, Specific cognitive domains, Lifestyle, Memory and learning, Stress, PET Imaging, Subjective complaints, Apathy, Fatigue, Depression, Neuropsychiatric Behaviors, Caregiver burden, Multisensory Stimulation, Not specified, Speed/accuracy of processing]	Memory and learning	Goal-setting and performance, Psychoeducation, Pleasant activities, Standardised cognitive task performance, Standardised skill performance, Conversation, N/A, Reality orientation, Everyday activity in context (e.g., banking), Reminiscence, Relaxation	Specific cognitive domains, Global cognitive function, Standardised skill performance, Goal-setting and performance, Standardised cognitive task performance, Memory and learning, Psychoeducation, Activities of daily living, Everyday activity in context (e.g., banking)
What was the statistical tests used?	[t-test, Analysis of variance, Repeated Measures Analysis of Variance, Linear Mixed Models, Generalised Estimating Equations, Regression, Mixed effects models, None, N/A, ‘‘Fisher s Test’’, Repeated Measures General Linear Model, Power calculation, Linear regression, Analysis of covariance, Mann–Whitney U-test, Pearson chi squared, Monte-Carlo significance, Jonckheere–Terpstra test, ‘‘Bonferroni s multiple comparison test’’, Benjamini–Hochberg procedure - FDR, Kruskal–Wallis test, Wilcoxon test, Kolmogorov–Smirnov, ANCOVA, Mediating effect model, chi-square test, ‘‘Pearson s correlation’’, Shapiro–Wilk test, constrained longitudinal data analysis (cLDA)]	t-test	Student’s t test for independent samples, Wilcoxon’s non-parametric test, Student’s t test for paired samples	Kolmogorov–Smirnov, Student’s t test for independent samples, Wilcoxon test, Student’s t test for paired samples

LLAMA 2 (Touvron et al., 2023) and Gemini (Team et al., 2023) on similar tasks.

In addition, it is also important to highlight potential challenges with the retriever component, which can impact the accuracy of information retrieval from documents and ultimately the generation, and should be considered for comprehensive improvement of QA framework. Our manual evaluation on 100 incorrect responses by GPT-4 with straightforward prompting demonstrated that the retrieved context was not relevant in 38% of the cases, which could impact the actual accuracy of the models on the CogTale dataset. This necessitates investigating different approaches for retrieval as part of future work. However, model’s incorrect responses in 62% of the cases for which the retrieved chunks contained relevant information highlights issues as stated above.

We summarize the overall findings below by including the strengths and weaknesses of the models:

Strengths:

- Single-Choice and Multiple-Choice Questions: These models excel in understanding the context provided in the question and selecting the most appropriate answer option from the given choices.
- Number Extraction: These models (particularly GPT-4) can identify and extract numerical information from text, facilitating accurate responses to questions involving numerical data.
- Document Answer Extraction: These models using the framework are adept at scanning documents to locate relevant information and extract answers, leveraging their reasoning over reliance on the knowledge base.
- Versatility: These models can handle different question types, including inferential and numeric questions, demonstrating versatility in addressing diverse user queries.
- Response Adaptability: They can adapt their response generation process based on the question category and input format, ensuring suitability for various QA tasks.

Weakness:

- Single-Choice and Multiple-Choice Questions: These models occasionally provide incorrect or incomplete answers due to limitations in their understanding of the context, or selection of more options than the actual number of answers (hallucination). GPT-3.5 was generally found to select more options than GPT-4, indicating GPT-4’s advanced reasoning capabilities to filter out many irrelevant options.
- Number Extraction: Despite their ability to identify numerical information, these models struggle to answer questions where a numerical value is required to be inferred.
- Document Answer Extraction: Inaccuracies or ambiguities in the document text can lead to errors or misunderstandings in answer extraction, impacting the reliability of responses. The models may still provide an answer even when the relevant information was not present in the extracted chunks or passages.
- Inference Challenges: They may struggle to interpret or infer responses when the answers are not directly stated, affecting the accuracy of responses, particularly in inferential questions.
- Response Variability: Occasionally, these models may provide responses outside the list of options.

In future work, it would also be valuable to explore how variations in the number of choices affect the performance of models on the single-choice and multiple-choice questions. Additionally, investigating how the complexity of number extraction questions impacts the accuracy could provide further insights.

Table 9
Selected Studies for the experiments from the Cogtale platform.

No.	Title of the Studies selected from the Cogtale platform
1	Benefits of Training Working Memory in Amnestic Mild Cognitive Impairment: Specific and Transfer Effects - (Carretti et al., 2013)
2	Cognitive training in older adults with Mild Cognitive Impairment: Impact on cognitive and functional performance - (Brum et al., 2009)
3	Effectiveness of a Visual Imagery Training Program to Improve Prospective Memory in Older Adults with and without Mild Cognitive Impairment: A Randomized Controlled Study - (Lajeunesse et al., 2022)
4	Toward rational use of cognitive training in those with mild cognitive impairment - (Hampstead et al., 2023)
5	Impact of metacognition and motivation on the efficacy of strategic memory training in older adults: Analysis of specific, transfer and maintenance effects - (Carretti et al., 2011)
6	Repetition-lag training to improve recollection memory in older people with amnestic mild cognitive impairment. A randomized controlled trial - (Finn and McDonald, 2015)
7	Effects of reality orientation therapy on elderly patients in the community - (Baldelli et al., 1993)
8	Efficacy of a cognitive intervention program in patients with mild cognitive impairment - (Rojas et al., 2013)
9	Computerized Structured Cognitive Training in Patients Affected by Early-Stage Alzheimer's Disease is Feasible and Effective: A Randomized Controlled Study - (Cavallo et al., 2016)
10	Efficacy of the Ubiquitous Spaced Retrievalbased Memory Advancement and Rehabilitation Training (USMART) program among patients with mild cognitive impairment: a randomized controlled crossover trial - (Han et al., 2017)
11	Cognitive rehabilitation combined with drug treatment in Alzheimer's disease patients: a pilot study - (Bottino et al., 2005)
12	Cognitive rehabilitation in patients with mild cognitive impairment - (Kurz et al., 2009)
13	The PACE Study: A Randomized Clinical Trial of Cognitive Activity Strategy Training for Older People with Mild Cognitive Impairment - (Vidovich et al., 2015)

Furthermore, it is essential to discuss the characteristics of the CogTale dataset and the potential difficulty of some questions. The dataset comprises generic questions intended to extract information from research papers (documents). However, due to their generic nature, the keywords or vocabulary in the questions may differ occasionally from those used in the papers. As a result, relevant (or most similar) context may not always be extracted, thus adding difficulty to such QA tasks.

Here, it is important to acknowledge the potential for ambiguity in the interpretation of text, especially for human readers who may lack domain expertise. The CogTale dataset, while valuable for its diverse range of questions, may still pose challenges in terms of subjective interpretation. Human readers, particularly those not well-versed in the specific domain, may encounter difficulties in discerning the most relevant information from research papers. This ambiguity highlights the inherent limitations in relying solely on language models for accurate information extraction, as disagreements in interpretation may arise. Therefore, to ensure the reliability and effectiveness of language models in QA tasks, it becomes imperative to strive for consensus among multiple domain experts. By incorporating the insights and expertise of diverse specialists, we can enhance the reliability of language model-generated responses and mitigate the impact of subjective interpretation.

Additionally, the dataset includes questions with options such as [No, N/A, Not Specified, Yes, ...] or [Yes - fully described, Yes - partially described, Yes, No, ...], where the model may select an option that is very close to the actual answer, such as 'Yes - fully described' and 'Yes' or 'No', 'N/A' and 'Not Specified' particularly when the context is not clear or missing. It is crucial to establish criteria for selecting such types of options, which is another aspect we plan to address in future work.

Finally, to run the same queries on these models, GPT-4 charged 26.54$ which is 15 times higher than that of GPT-3.5-turbo (1.77$). Therefore, while GPT-4 exhibits superior performance, the financial burden associated with its usage must be carefully weighed against the incremental gains in accuracy, especially in scenarios where cost-effectiveness is a paramount consideration. This economic dimension adds a layer of complexity to the decision-making process when selecting a model for practical applications.

6. Conclusion

In conclusion, this paper's evaluation of language models (GPT-3.5 and GPT-4) in question-answering tasks based on RAG, facilitated by the CogTale dataset encompassing trial information and diverse inquiries and different question formats, has provided a nuanced perspective on their performance on information retrieval-based QA task. While both models demonstrated proficiency in various question types, they encountered challenges in tasks requiring inference from contextual cues. GPT-4 exhibited notable proficiency in certain question categories, adeptly grasping contextual cues to deliver coherent responses. However, the study also unveiled vulnerabilities when answering questions with multiple choices and number extraction. Additionally, the retriever plays an important role in such QA tasks and requires careful selection and evaluation of the retrieval approaches. Our work highlights areas for further exploration and improvements and development of robust strategies to improve models performance on these task. As language models continue to evolve, this evaluation serves as a guiding compass, navigating us toward more refined and versatile models that transcend the boundaries of textual and numerical comprehension, ultimately advancing the landscape of natural language processing.

7. Threat to validity

As we move towards bridging the gap in evaluating LLMs on information retrieval QA task across diverse question formats using the CogTale dataset, it is imperative to acknowledge potential threats to the validity of our study's outcomes. The language models exhibit dynamic behavior, and a model update has the potential to enhance or diminish its performance on the dataset. Thus, the results on CogTale dataset may improve on a newer version of the GPT-4.

CRediT authorship contribution statement

Zafaryab Rasool: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Stefanus Kurniawan:** Validation, Software, Formal analysis, Data curation. **Sherwin Balugo:** Software, Data curation.

Scott Barnett: Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **Rajesh Vasa:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **Courtney Chesser:** Writing – review & editing, Data curation. **Benjamin M. Hampstead:** Data curation. **Sylvie Belleville:** Data curation. **Kon Mouzakis:** Writing – review & editing, Resources, Project administration, Conceptualization. **Alex Bahar-Fuchs:** Writing – review & editing, Resources, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank the reviewers for their insightful and constructive comments, which have significantly enriched this work.

Appendix. Studies used

The titles of the studies used in the analysis for this work are presented in Table 9.

References

Acharya, A., Singh, B., Onoe, N., 2023. LLM based generation of item-description for recommendation system. In: Proceedings of the 17th ACM Conference on Recommender Systems. pp. 1204–1207.

Aher, G.V., Arriaga, R.I., Kalai, A.T., 2023. Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning. PMLR, pp. 337–371.

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., Zhang, J., Li, J., Hou, L., 2023. Benchmarking foundation models with language-model-as-an-examiner. arXiv:2306.04181.

Baldelli, M., Pirani, A., Motta, M., Abati, E., Mariani, E., Manzi, V., 1993. Effects of reality orientation therapy on elderly patients in the community. Arch. Gerontol. Geriatrics 17 (3), 211–218.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P., 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv:2302.04023.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., Abdelrazek, M., 2024. Seven failure points when engineering a retrieval augmented generation system. arXiv preprint arXiv:2401.05856.

Bian, N., Han, X., Sun, L., Lin, H., Lu, Y., He, B., 2023. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. arXiv:2303.16421.

Bottino, C.M., Carvalho, I.A., Alvarez, A.M.M., Avila, R., Zukauskas, P.R., Bustamante, S.E., Andrade, F.C., Hototian, S.R., Saffi, F., Camargo, C.H., 2005. Cognitive rehabilitation combined with drug treatment in Alzheimer's disease patients: A pilot study. Clinical Rehabil. 19 (8), 861–869.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Brum, P.S., Forlenza, O.V., Yassuda, M.S., 2009. Cognitive training in older adults with mild cognitive impairment: Impact on cognitive and functional performance. Dementia Neuropsychol. 3, 124–131.

Carretti, B., Borella, E., Fostinelli, S., Zavagnin, M., 2013. Benefits of training working memory in amnestic mild cognitive impairment: Specific and transfer effects. Int. Psychogeriatr. 25 (4), 617–626.

Carretti, B., Borella, E., Zavagnin, M., De Beni, R., 2011. Impact of metacognition and motivation on the efficacy of strategic memory training in older adults: Analysis of specific, transfer and maintenance effects. Arch. Gerontol. Geriatrics 52 (3), e192–e197.

Cavallo, M., Hunter, E.M., van der Hiele, K., Angilletta, C., 2016. Computerized structured cognitive training in patients affected by early-stage Alzheimer's disease is feasible and effective: A randomized controlled study. Arch. Clin. Neuropsychol. 31 (8), 868–876.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X., 2023. A survey on evaluation of large language models. arXiv:2307.03109.

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N.A., Gardner, M., 2021. A dataset of information-seeking questions and answers anchored in research papers. arXiv preprint arXiv:2105.03011.

Espejel, J.L., Ettifouri, E.H., Alassan, M.S.Y., Chouham, E.M., Dahhane, W., 2023. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. Nat. Lang. Process. J. 5, 100032.

Ferguson, J., Gardner, M., Hajishirzi, H., Khot, T., Dasigi, P., 2020. IIRC: A dataset of incomplete information reading comprehension questions. arXiv preprint arXiv:2011.07127.

Finn, M., McDonald, S., 2015. Repetition-lag training to improve recollection memory in older people with amnestic mild cognitive impairment. A randomized controlled trial. Aging, Neuropsychol. Cognit. 22 (2), 244–258.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., Berant, J., 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. Trans. Assoc. Comput. Linguist. 9, 346–361.

Hampstead, B.M., Stringer, A.Y., Iordan, A.D., Ploutz-Snyder, R., Sathian, K., 2023. Toward rational use of cognitive training in those with mild cognitive impairment. Alzheimer's Dementia 19 (3), 933–945.

Han, J.W., Son, K.L., Byun, H.J., Ko, J.W., Kim, K., Hong, J.W., Kim, T.H., Kim, K.W., 2017. Efficacy of the ubiquitous spaced retrieval-based memory advancement and rehabilitation training (USMART) program among patients with mild cognitive impairment: A randomized controlled crossover trial. Alzheimer's Res. Therapy 9 (1), 1–8.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X., 2019. PubMedQA: A dataset for biomedical research question answering. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, pp. 2567–2577. http://dx.doi.org/10.18653/v1/D19-1259, URL https://aclanthology.org/D19-1259.

Johnson, J., Douze, M., Jégou, H., 2019. Billion-scale similarity search with GPUs. IEEE Trans. Big Data 7 (3), 535–547.

Kalyan, K.S., 2023. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat. Lang. Process. J. 100048.

Kamalloo, E., Dziri, N., Clarke, C.L., Rafiei, D., 2023. Evaluating open-domain question answering in the era of large language models. arXiv preprint arXiv:2305.06984.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. Adv. Neural Inf. Process. Syst. 35, 22199–22213.

Krithara, A., Nentidis, A., Bougiatiotis, K., Paliouras, G., 2023. BioASQ-QA: A manually curated corpus for biomedical question answering. Sci. Data 10 (1), 170.

Kurz, A., Pohl, C., Ramsenthaler, M., Sorg, C., 2009. Cognitive rehabilitation in patients with mild cognitive impairment. Int. J. Geriatric Psychiatry: J. Psychiatry Late Life Allied Sci. 24 (2), 163–168.

Lajeunesse, A., Potvin, M.J., Labelle, V., Chasles, M.J., Kergoat, M.J., Villalpando, J.M., Joubert, S., Rouleau, I., 2022. Effectiveness of a visual imagery training program to improve prospective memory in older adults with and without mild cognitive impairment: A randomized controlled study. Neuropsychol. Rehabil. 32 (7), 1576–1604.

Lee, G.-G., Latif, E., Wu, X., Liu, N., Zhai, X., 2024. Applying large language models and chain-of-thought for automatic scoring. Comput. Educ.: Artif. Intell. 100213.

Levine, Y., Ram, O., Jannai, D., Lenz, B., Shalev-Shwartz, S., Shashua, A., Leyton-Brown, K., Shoham, Y., 2022. Huge frozen language models as readers for open-domain question answering. In: ICML 2022 Workshop on Knowledge Retrieval and Language Models. URL https://openreview.net/forum?id=z3Bxu8xNJaF.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474.

Liu, C., Li, X., Shang, L., Jiang, X., Liu, Q., Lam, E., Wong, N., 2023. Gradually excavating external knowledge for implicit complex question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 14405–14417.

OpenAI, 2023. GPT-4 technical report. arXiv:2303.08774.

Paliouras, G., Krithara, A., 2014. A challenge on large-scale biomedical semantic indexing and question answering.

Pereira, J., Fidalgo, R., Lotufo, R., Nogueira, R., 2023. Visconde: Multi-document QA with GPT-3 and neural reranking. In: European Conference on Information Retrieval. Springer, pp. 534–543.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D., 2023. Is ChatGPT a general-purpose natural language processing task solver? arXiv:2302.06476.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y., 2023. In-context retrieval-augmented language models. arXiv preprint arXiv:2302.00083.

Rasool, Z., Barnett, S., Willie, D., Kurniawan, S., Balugo, S., Thudumu, S., Abdelrazek, M., 2024. LLMs for test input generation for semantic caches. arXiv preprint arXiv:2401.08138.

Rojas, G.J., Villar, V., Iturry, M., Harris, P., Serrano, C.M., Herrera, J.A., Allegri, R.F., 2013. Efficacy of a cognitive intervention program in patients with mild cognitive impairment. Int. Psychogeriatr. 25 (5), 825–831.

Sabates, J., Belleville, S., Castellani, M., Dwolatzky, T., Hampstead, B.M., Lampit, A., Simon, S., Anstey, K., Goodenough, B., Mancuso, S., et al., 2021. CogTale: An online platform for the evaluation, synthesis, and dissemination of evidence from cognitive interventions studies. Syst. Rev. 10 (1), 1–11.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., tau Yih, W., 2023. REPLUG: Retrieval-augmented black-box language models. arXiv: 2301.12652.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al., 2023. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al., 2023. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Vidovich, M.R., Lautenschlager, N.T., Flicker, L., Clare, L., McCaul, K., Almeida, O.P., 2015. The PACE study: A randomized clinical trial of cognitive activity strategy training for older people with mild cognitive impairment. Am. J. Geriatric Psychiatry 23 (4), 360–372.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R., 2023. A survey of large language models. arXiv:2303.18223.