A SURVEY OF EVALUATING QUESTION ANSWERING TECHNIQUES IN THE ERA OF LARGE LANGUAGE MODEL LLM

Jassir Altheyabi, Khaled Almuteb and Bader Alshemaimry

College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

Large language models (LLMs) are increasingly popular in academia and industry due to their exceptional performance in various applications. As LLMs play a crucial role in research and everyday use, their evaluation becomes essential, not only at the task level but also at the societal level, to understand potential risks. This article provides a comprehensive review of LLM evaluation methods, focusing on three key dimensions: what to evaluate, where to evaluate, and how to evaluate. It covers evaluation tasks in areas such as natural language processing, reasoning, medical usage, ethics, education, natural and social sciences, and agent applications. The article also discusses evaluation methods and benchmarks, addressing where and how to assess LLM performance. Additionally, it summarizes instances of success and failure of LLMs across different tasks and highlights important aspects to consider in the evaluation process.

KEYWORDS

Question answering techniques, Large Language Model, Knowledge base question answering, open domain questions answering.

1. BACKGROUND

The emergence of large language models (LLMs) has revolutionized natural language processing (NLP), particularly in the domain of question answering (QA). LLMs, trained on massive amounts of text data, possess the remarkable ability to generate human-quality responses to a wide range of prompts and questions. This inherent capability makes them well-suited for tackling QA tasks, where the objective is to provide accurate and informative answers to user queries, Question answering is a crucial technology in the field of human-computer interaction, and it has found wide application in scenarios like search engines, intelligent customer service, and QA systems. The measurement of accuracy and efficiency in QA models will have significant implications for these applications. According to Liang et al. [28], among all the evaluated models, Instruct GPT davinci v2 (175B) exhibited the highest performance in terms of ac- curacy, robustness and fairness across the 9 QA scenarios. Both GPT-3.5 and ChatGPT demonstrate significant advancements compared to GPT-3 in their ability to answer general knowledge questions. In most domains, ChatGPT surpasses GPT-3.5 by more than 2 per- cent in terms of performance, In recent years, the evaluation of LLM performance in QA tasks has gained significant traction. However, this endeavour is not without its challenges. Traditional QA evaluation metrics, such as accuracy and F1 score, may not adequately capture the nuances of LLM performance. This stems from the fact that LLMs can occasionally produce factually correct responses that lack relevance to the user's query. Additionally, LLMs often generate

David C. Wyld et al. (Eds): NLAI, IBCOM, MLDS, GridCom, SCAI, ICCSEA, SPPR – 2024
pp. 53-73, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142205

multiple plausible responses for a given query, making it challenging to assess their effectiveness conclusively.

2. INTRODUCTION

Large language models (LLMs) have emerged as powerful tools for natural language pro- cessing (NLP) tasks, including question answering (QA). LLMs are trained on massive amounts of text data and can generate human-quality text in response to a wide range of prompts and questions. This makes them well-suited for QA tasks, where the goal is to provide accurate and informative answers to user queries. In recent years, there has been a growing interest in evaluating the performance of LLMs on QA tasks. However, this task is not without its challenges. Traditional QA evaluation metrics, such as accuracy and F1 score, may not be well-suited for evaluating LLMs. This is because LLMs can sometimes generate responses that are factually correct but are not relevant to the user's query. Additionally, LLMs can often generate multiple possible responses to a given query, which can make it difficult to assess their performance. In this paper, we survey a variety of methods for evaluating question answering techniques on LLMs. We discuss the strengths and weaknesses of each method and provide recommendations for choosing the most appropriate method for a given task. The advent of large language models (LLMs) has revolutionized the landscape of natural language processing (NLP), ushering in a new era of capabilities and challenges. Among these advancements, question answering (QA) has emerged as a critical area of focus, with LLMs demonstrating remarkable ability to comprehend and respond to user queries. How- ever, evaluating the performance of these models on QA tasks presents a unique set of complexities. Traditional QA evaluation metrics, such as accuracy and F1 score, while valuable, may not fully capture the nuances of LLM-driven QA. These metrics often focus solely on factuality, overlooking the crucial aspect of response relevance to the user's intent. Additionally, LLMs' propensity to generate multiple potential responses complicates the assessment process. To address these challenges, this paper delves into a comprehensive survey of evaluation methods tailored to assess the effectiveness of question answering techniques on LLMs. We explore a range of approaches, encompassing both quantitative and qualitative measures, considering their strengths, limitations, and suitability for specific evaluation scenarios.

2.1. Objectives of the Research

in this survey, our objectives are to review research papers to understand the capabilities of LLMs in question answering and existing evaluation metrics and methodologies for LLM question answering and identify the challenges and limitations of current evaluation meth- ods, also to review the current techniques used by researchers for evaluating LLM question answering performance.

2.2. Significance of Question Answering Techniques on LLMs

Question-answering (QA) techniques are rapidly gaining significance in the realm of large language models (LLMs), particularly in the context of natural language processing (NLP). As LLMs continue to evolve and become more sophisticated, their ability to effectively answer questions holds immense potential for a wide range of applications. Enhancing Human-Computer Interaction, LLMs equipped with advanced Question answering capabilities can revolutionize human-computer interaction by enabling more natural and intuitive communication. Instead of relying on rigid keyword-based search mechanisms, users can engage in open-ended conversations with AI systems, asking questions directly and receiving comprehensive and informative responses. This shift towards conversational AI will significantly enhance the user

experience and make AI more accessible to a broader audience. also Powering Intelligent Search Engines leads to the ability of LLMs to answer questions accurately and comprehensively makes them ideal candidates for powering intelligent search engines. Unlike traditional search engines that simply return a list of relevant documents, LLM-based search engines can directly provide answers to user queries, saving time and effort. This approach is particularly beneficial for complex or open-ended questions that require more than just a list of links. For the Driving Knowledge Discovery and Summarization, LLMs can play a crucial role in knowledge discovery and summarization tasks by extracting and analyzing information from vast amounts of text data. By understanding the context and relationships between concepts, LLMs can generate concise and informative summaries of complex topics, making it easier for users to grasp key information quickly and efficiently. In the Fostering Creative Writing and Storytelling, LLMs can be employed to assist with creative writing and storytelling tasks by generating new ideas, developing plot structures, and crafting engaging narratives. Their ability to understand and manipulate language enables them to produce creative content that is both original and captivating. From Addressing Ethical Considerations and Potential Biases perspective, LLM-based QA techniques continue to advance, it is crucial to carefully consider ethical implications and potential biases. Ensuring that LLMs are trained on diverse and unbiased data sets is essential to prevent the perpetuation of harmful stereotypes or discriminatory practices.

3. METHODOLOGY

Papers in this format must not exceed twenty (20) pages in length. Papers should be submitted to the secretary AIRCC. Papers for initial consideration may be submitted in either .doc or .pdf format. Final, camera-ready versions should take into account referees' suggested amendments.

3.1. Research Methodology

The main goal of this paper is to explore question answering to validate the effectiveness of its capabilities in the era of the large language model (LLMs). I highlighted the methods conducted in question answering using LLMs by different approaches and techniques.

The methodology steps in the literature review are as follows:

1. Define the research questions.
2. Carry out the review (search, list, and evaluate primary studies, extract, and synthesize data to produce a concrete result).
3. Evaluate the papers.
4. Report the results.

3.2. Research Questions

RQ1. To what extent do different LLM architectures (e.g., transformer-based vs. recurrent neural networks) and training objectives impact their performance on specific QA tasks?
RQ2. How do factors like data size, training time, and model size influence the effectiveness of LLMs in QA compared to traditional techniques?
RQ3. What are the unique strengths and weaknesses of LLMs in answering complex, open-ended, or challenging questions compared to traditional methods?
RQ4. What are the potential biases and limitations of LLMs in QA, and how can they be mitigated or addressed through evaluation methods?

3.3. Search Criteria

As a preliminary step and before starting to look for research papers, a list of papers must be carefully chosen to improve the possibility of obtaining the most broad and relevant resources. In the literature review, the following criteria are used to select the source papers: The database must include journals and conference proceedings that cover question-answering.

3.4. Search Strings

We determine the keywords that should be used to find any articles that can contribute to the literature review as follows:

((Question answering techniques" OR" Large Language Model" OR" LLM" OR "Knowledge base question answering" OR " open domain questions answering") AND ("Evaluation" OR "benchmark" OR "Performance")) techniques, Knowledge base question answering, large language models evaluation. To minimize the redundancy of journals and proceedings across databases, the list of papers is reduced where possible.

3.5. Inclusion and Exclusion Criteria

Table 1: Inclusion and Exclusion Criteria

Exclusion Criteria	Inclusion Criteria
Papers that developed for specific purposes Magazines, Posters, short papers, and tutorials Papers that are not directly related to question answering systems Papers that are written other than the English language Papers that are insufficient in academic research	Papers that address large language models, techniques, approaches, and methods Papers that focused on measuring question answering techniques Papers that are mainly about LLMs Paper that is published after 2019 Papers that are published in high ranking journals

3.6. Quality Assessment Criteria

So, I examined the remaining papers against quality assessment criteria as shown in Table 2.

Table 2: Quality Assessment Criteria

Sr. No	Quality Assessment Criteria
1	Is the paper based on a strong research methodology and is not just a report promising experimental evaluation?
2	Does the paper present clear goals of the research?
3	Does the paper present an appropriate description of the paper content?
4	Does the research methodology describe well-designed steps that navigate the main contribution of the research?
5	Is the research methodology sound and appropriate in terms of the main contribution of the research?
6	Is the collected data (if any) commonly used by researchers?
7	Does the paper build a clear result and lessons learned?

3.7. The Primary Selections

Then, based on the literature review dataset, I finalized the data collection process and summarized below:

Table 3: Publication Setting

Publication setting	Count	Percentage
Academic	25	78%
Practitioner	7	22%
Mixed	0	0%
Other	0	0%

Table 4: Paper Methodology

Paper Methodology	Count	Percentage
Theoretical	11	34%
Experimental	21	66%
Both	0	44%
Other	0	0%

4. RELATED WORK

Previous studies on question answering with LLMs (Language Model-based Models) have explored the use of these models to improve the accuracy and effectiveness of question-answering systems. LLMs are powerful machine learning models that have been trained on large amounts of text data, enabling them to understand and generate human-like language, One notable study in this area is the work by Radford et al. [33], who introduced the concept of the OpenAI GPT (Generative Pre-trained Transformer). The authors demonstrated that fine-tuning the GPT model on specific question-answering tasks can lead to impressive results, surpassing previous state-of-the-art methods. They showed that by using a com- bination of unsupervised pre-training and supervised fine-tuning, the model can effectively answer questions given a passage of text. Another important study is by Liu et al. [24], who proposed a novel approach called BERT (Bidirectional Encoder Representations from Transformers). BERT is a pre-trained language model that has achieved remarkable success in various natural language processing tasks, including question-answering. The authors showed that by fine-tuning BERT on a specific question-answering dataset, it outperformed existing models, demonstrating its effectiveness in understanding and generating accurate answers, Furthermore, Chen et al. [11] conducted research on incorporating external knowledge into question-answering systems using LLMs. They proposed a method called Knowledge-Infused LSTM (KI-LSTM), which integrates external knowledge sources into the LSTM-based model for better question understanding and answer generation. The study demonstrated that leveraging external knowledge can significantly improve the performance of question-answering systems. In addition, Joshi et al. [18] explored the use of LLMs for multi-hop question answering, where answering a question requires reasoning over multiple pieces of information. They introduced a dataset called HotpotQA, which consists of questions that require multi-hop reasoning to answer accurately. The authors showed that by fine-tuning BERT on this dataset, the model achieved state-of-the-art performance on multi-hop question-answering tasks, Lastly, Min et al. [?]investigated the use of LLMs for open-domain question answering, where questions can be asked about any topic. They proposed a method called OpenQA, which uses the GPT model to generate answers given a question and a large

collection of documents. The study demonstrated that OpenQA outperformed existing methods in terms of both accuracy and efficiency.

Table 5: Summary of large language models (LLMs) studies

Study	Authors	Publication year	Source type
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"	Jacob Devlin, et al.	2019	Academic Journal
"Improving Language Understanding by Generative Pre-Training"	Alec Radford, et al.	2018	Academic Journal
"Language Models are Unsupervised Multitask Learners"	Alec Radford, et al.	2019	Academic Journal
"BERTology Meets Biology: Interpreting Attention in Protein Language Models"	Guangyu Zheng, et al.	2020	Academic Journal
"OpenAI GPT"	Alec Radford, et al.	2019	Academic Journal
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"	Devlin, Jacob et al.	2018	Academic Journal
"RoBERTa: A Robustly Optimized BERT Pretraining Approach"	Liu, Yinhan et al.	2019	Academic Journal
"XLNet: Generalized Autoregressive Pretraining for Language Understanding"	Yang, Zhilin et al.	2019	Academic Journal
"ALBERT: A Lite BERT for Self-supervised Learning of Language Representations"	Lan, Zhenzhong et al.	2020	Academic Journal
"T5: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"	Raffel, Colin et al.	2019	Academic Journal

5. LITERATURE REVIEW

5.1. Overview of Large Language Models (LLMs)

Large Language Models (LLMs) are a type of artificial intelligence system that has the ability to understand and generate human language. These models are designed to process and analyze vast amounts of text data, allowing them to learn patterns, relationships, and structures within language. LLMs have gained significant attention and popularity in recent years due to their remarkable ability to generate coherent and contextually relevant text, And One of the key features of LLMs is their capacity to generate text that closely resembles human language. These models are trained on a wide range of textual sources, such as books, articles, and websites, allowing them to capture the nuances and complexities of language usage. By leveraging this training data, LLMs can generate text that is grammatically correct, contextually appropriate, and coherent.

The architecture of LLMs typically consists of multiple layers of neural networks. These networks are trained using a technique called unsupervised learning, where the model learns from the input data without explicit guidance or labelling. LLMs employ deep learning algorithms, such as transformers, which enable them to process and understand large chunks of text data efficiently, Also LLMs have numerous applications across various domains. They are widely used in natural language processing tasks such as machine translation, sentiment analysis, question answering systems, and chatbots. Additionally, LLMs have been utilized in creative tasks like generating poetry, composing music, and writing stories.

However, LLMs also face challenges and ethical concerns. One significant concern is the potential for bias in generated text due to the biases present in the training data. Addition- ally, there are concerns about the misuse of LLMs for spreading misinformation or generating malicious content, Large Language Models (LLMs) have revolutionized the field of natural language processing by their ability to understand and generate human language. These models have numerous applications but also come with ethical considerations that need to be addressed.

5.2. Evolution of Question-Answering Techniques

On the Common senseQA and Social IQA benchmarks, ChatGPT performs somewhat worse than GPT-3.5. This is explained by ChatGPT's cautious approach, which makes it more likely to refuse to respond when there is not enough information. Models with supervised fine-tuning, like ChatGPT and V´ıcuna, perform exceptionally well with almost flawless scores, outperforming unsupervised models by a wide margin [2, 3]. The usefulness of ChatGPT was assessed by Laskar et al. [16] using a variety of academic datasets and tasks, including responding to queries, summarizing text, writing code, applying common sense reasoning, resolving mathematical puzzles, translating languages, identifying bias, and handling ethical dilemmas. All things considered, LLMs demonstrate perfect performance on QA tests and have the capacity to further improve their competency in social, event, and temporal commonsense knowledge. later on. There are further generating chores to investigate. Pu and Demberg [30]showed that ChatGPT outperforms the earlier SOTA supervised model in the field of phrase style transfer by training on the same subset for few-shot learning, as seen by the higher BLEU score. However, ChatGPT's performance still devi- ates greatly from human behaviour when it comes to managing sentence structure formality. Chia et al. [13] found that LLMs consistently do well in writing assignments in a variety of categories, including professional, informational, argumentative, and creative writing. This result suggests that LLMs have a general level of writing proficiency. Chen et al. [12]found that ChatGPT is highly effective at evaluating text quality from many perspectives, even when there isn't of reference texts, outperforming the majority of automated metrics already in use. Out of all the testing techniques examined, using ChatGPT to provide numerical scores for text quality turned out to be the most dependable and efficient strategy.

Yopeng Chang et al. [10] evaluates the performance of GPT-3.5, GPT-4, and BARD models on different reasoning tasks in a zero-shot setting. The authors conduct a thorough technical evaluation on eleven distinct datasets to assess the reasoning capacity of these models.

.and the study provides empirical evidence that ChatGPT-4 outperforms both ChatGPT-3.5 and BARD in the zero-shot setting across most evaluated tasks .The superiority of GPT-4 over GPT-3.5 can be attributed to its larger size and NLP efficiency, but this is not evi- dent for BARD .The three models show limited proficiency in Inductive, Mathematical, and Multi-hop Reasoning Tasks .he also propose a set of engineered prompts that enhance the zero-shot setting performance of all three models .The HotpotQA dataset, which contains over 100,000 question-answer pairs derived from Wikipedia, is used to evaluate multi-hop reasoning .The paper also mentions prompts proposed by Kojima and Shane Gu [20] for mathematical problems and

prompt engineering proposed by Bang et al.[6] for judging plausibility ., the study evaluates the reasoning ability of LLMs and presents findings on their performance in a zero-shot setting, highlighting the superiority of ChatGPT-4 and the limitations of all three models in certain reasoning tasks.

Yuchen Zhuang et al, [41] The paper introduces a dataset called ToolQA, which evaluates the ability of Large Language Models (LLMs) to use external tools for question answering, The ToolQA dataset evaluates LLMs external tool use reasoning abilities and finds that there exists a marginally elevated incidence of hallucination and errors of lengthy context whilst responding to arduous inquiries. This phenomenon can be ascribed to the intricacy inherent in challenging questions, which frequently necessitate the formulation of additional resources to address the given queries, Linyong Nan et al. [29] presents a novel database that is characterized by a long-form structure. This database is specifically created for the purpose of assessing the manner in which Large Language Models (LLMs) engage with a SQL interpreter. The assigned task requires LLMs to employ a strategic approach in generating numerous SQL queries to obtain an adequate amount of data from the database. Furthermore, a multi-agent evaluation framework is proposed in this study to replicate the academic peer-review process. The objective of this framework is to improve the accuracy and dependability of the researcher's assessments, The authors were required through their evaluation, to suggest the expansion of the dataset to augment the statistical significance of the findings. Furthermore, his current approach lacks a comprehensive human evaluation to determine the correlation between his automated evaluation techniques and human judgment. Additionally, the authors neglected to investigate the performance of LLM agents when integrated with external modules on analogous tasks that have less demanding criteria.

Catherine Kosten et al. [21] introduce a newly developed SPARQL benchmark dataset called Spider4SPARQL, which has presented a collection of intricate and unprecedented SPARQL queries that encompass a wide range of complexities. Notably, these queries en- compass 138 distinct domains. The complexity inherent in this benchmark dataset offers a unique opportunity to assess the capabilities and limitations of contemporary KGQA systems. In this regard, the authors of this dataset conduct a comprehensive evaluation by comparing their system to state-of-the-art KGQA systems and LLMs. Of noteworthy mention is the fact that the LLMs achieve a mere 45% execution accuracy, thereby showcasing the formidable nature of Spider4SPARQL as a benchmark dataset that will stimulate future research endeavors, [40] notices a notable deficiency in the current landscape pertains to the absence of a practical standard for assessing the efficacy of grounding Language and Knowledge Models (LLMs) on heterogeneous sources of knowledge. Moreover, there exist two primary obstacles that need to be addressed. The first challenge involves resolving Two-hop multi-source questions, which necessitate the retrieval of information from both open-domain structured and unstructured knowledge sources. In this regard, it is crucial to underscore the importance of retrieving information from structured knowledge sources to accurately answer the aforementioned questions. The second challenge pertains to the generation of symbolic queries, particularly SPARQL for Wikidata, which introduces an additional layer of complexity, the authors also introduce a novel approach that leverages multiple retrieval tools, including text passage retrieval and symbolic language-assisted retrieval. Our model outperforms previous approaches by a significant margin, demonstrating its effectiveness in addressing the above-mentioned reasoning challenges.

[21] this paper offers a comprehensive examination of different open-domain question-answering (QA) models, such as Language Models (LLMs), through manual assessment of their responses on a subset of NQ-OPEN, a widely recognized benchmark. The findings un- veil that although the true effectiveness of all models is greatly undervalued, the Instruct GPT (zero-shot) LLM experiences a notable increase of nearly +60% in performance.

6. DISCUSSION

6.1. Comparative Analysis of Question answering techniques

The following table shows Table 2. the main points between different techniques in question answering using LLMs :

No.	Author	Methodology	Limitation	Strength
1	Nan Liny- ong et al.	- Designing experiments to eval- uate different types of LLM agents. - Using a meta-review process to balance precision and recall in evaluation.	- Small evaluation - dataset due to bud- get constraints. - Lack of rigorous hu- man evaluation for alignment with auto- matic evaluation. - Limited explo- ration of LLM agents with external modules.	- Augmenting LLMs with a symbolic mod- ule - a SQL code interpreter. - Assessing LLM performance
2	Cather- ine Kosten et al.	- Pattern-based SPARQL query generation approaches - Crowdsourcing for natural language question generation - Rule-based paraphrasing for natural language question gen- eration NL question templates for natural language question generation	- - Pattern-based gen- eration approaches do not generalize well to vague and diverse questions.- Existing systems achieve only up to 45% execution accuracy. - Existing systems need substantial improve- ments to achieve higher execution accuracies.	- Analysis of the current state of NLS-PARQL benchmarks for KGQA tasks. - Introduction of Spider4SPARQL, a complex cross-domain benchmark dataset for KGQA tasks. - The dataset features 4,721 unique and novel SPARQL pairs and 166 multi-domain - knowledge graphs and ontologies.
3	Wenting Zhao et al.	- Retrieval augmented LLMs - Multiple retrieval tools includ- ing text passage retrieval and symbolic language-assisted re- trieval Generation of symbolic queries (e.g., SPARQL for Wikidata)	- Retrieval tool usage in each step needs optimization and trustworthiness. - Impact of extended prompts on retrieval- augmented language models. - Susceptibility to recency bias and document reordering strategies.	- Existing QA bench- marks are limited in evaluating retrieval- augmented language models. - The paper intro- duces a comprehensive dataset for evaluating grounding LLMs on heterogeneous knowl- edge sources. - The dataset includes two-hop multi-source questions and symbolic. - query generation.

4	Ehsan Kamalloo et al.	- Lexical matching- Manual evaluation - Regex matching - Automated evaluation models - Human evaluation	- Focus is limited to factoid information- seeking questions with short answers. - Lexical matching fails for more com- plex forms of QA. - Lexical matching struggles with gen- erative models and longer candidate answers.	- Evaluation method for open-domain QA models using lexical matching. - Linguistic analysis of failure cases in lexical matching. - Use of regular expressions for evaluation. - Inability of semantic matching to accurately evaluate LLMs.
5	Zafaryab Rasool et al.	- The paper describes the Cogtale dataset and the framework for evaluating LLM performance on QA tasks.	LLMs' performance diminishes on multiple-choice and number extraction questions. - LLMs may not be reliable for precise information ex- traction from documents. - Potential threats to validity include model updates and prompting - techniques.	- Investigating the performance of LLMs in information retrieval tasks. - Evaluating LLMs' performance on exact answer selection and numerical extraction. - Finding that LLMs perform well on single-choice and yes-no questions.
6	Ella Rabinovich et al.	Manual creation of a benchmark dataset with high - quality paraphrases for factual questions - Combining semantic consistency metric with additional measurements for building and evaluating a framework for factual QA reference-less performance prediction	- Semantic consis- tency measurement is limited to factual QA task. - The framework requires external knowledge" which may not always be available.	- Introducing and releasing a large extension of the PopQA dataset (PopQA-TP) with high-quality paraphrases. - Developing a prototype model for QA performance prediction.

7	Tal Schu s- ter	- Introducing the task of semi-extractive multi-source QA (SEMQA) - Generating a summarized and well-grounded answer by com- bining information from multiple sources - Extracting factual spans and connecting them with free-text connectors	- Limited to English questions, answers, and Wikipedia articles as sources. - Scope limited to questions in NQ and PAQ collections. - Semi-extractive for- mat not immune to model hallucinations or out-of-context is- sues.	- Introducing the task of semi-extractive multi-source QA (SEMQA) for answering multi-answer questions. - Creating a dataset of questions, relevant passages, and human- written semi-extractive answers. - Developing annotation pipeline for writers to compile answers using a quoting mechanism.
8	Tal Schus- ter	- Regularization of hierarchical tables - Table retrieval Prediction of math expres- sions, programs, or code for nu- merical reasoning	- The paper has some overlap with the functionality of OpenRT. - The focus of the paper is on table question answering tasks and LLM- based methods. - The authors plan to expand the methods included in the toolkit in the future.	- Development of a comprehensive toolkit for TableQA. - Introduction of a challenging LLM TableQA bench- mark. - Unification of different datasets under a single interface. - Support for multi- type tables and multi- modal data.

9	Fangkai Yang et al.	- Proposed model interaction paradigm for empowering LLMs with domain-specific knowledge. - Introduced MSQA dataset tie- lored for cloud domain QA. - Used lexical-overlap-based metrics and semantic. - Overlap-based metrics for evaluation.	- Dataset confirmed. to Microsoft Azure, impacting generaliz- ability in other do- mains. - Difficulty in set- ting the number of epochs properly in instruction tuning. - Lack of well-defined and automated metrics to evaluate LFQA. 019	- Introduction of MSQA dataset for evaluating LLMs in cloud domain QA. - Empowering LLMs with domain-specific knowledge for accurate answers in industrial scenarios.
10	Zis- han Guo et al.	- Natural language inference (NLI) based methods. - Question answering (QA) and generation (QG) based methods	- LLMs could suffer from private data leaks. - LLMs could yield inappropriate, harm- ful, or misleading content. - Concerns about the potential emergence of superintelligent systems without safeguards.	- Evaluation framework for information extrac- tion in LLMs. - Addressing safety is- sues and risks related to LLMs. - Systematic literature review of LLM evalua- tion efforts. - Highlighting the need for comprehensive eval- uation to address criti- cal issues.
11	Shahriar Golchin et al.	- The paper proposes the ”Data Contamination Quiz” as a method to detect data con- tamination in large language models.- The quiz involves pre- senting the model with four op- tions, one being the original in- stance and the others being per- turbed versions.	- LLMs exhibit inherent biases towards certain positions. - Positional biases can distort the as- sumption of random choice. - The likelihood of canonically ordered benchmark datasets may not exclusively reflect data contam- - ination.	- The paper proposes - the Data Contamina- tion Quiz as a method to detect and estimate data contamination in LLMs.- The quiz consists of a four- option multiple- choice format with perturbed versions of the original instance.- The method does not require access to the pre-training data of the LLMs.

12	Ning Bian et al.	- Conducted a series of experiments to evaluate ChatGPT's commonsense abilities.	- GPTs struggle with certain types of knowledge, including social and temporal commonsense. - GPTs contain misleading and overgeneralized commonsense knowledge.	- Investigated the commonsense abilities of large language models - Found that GPTs can achieve good accuracy in commonsense QA tasks
13	Yejin Bang et al.	- Evaluate the multitask, multilingual, and multimodal aspects of ChatGPT based on these datasets and a newly designed multimodal dataset.	- Limited number of samples for evaluation (30-200) - Recent updates related to safety concerns may not affect evaluation tasks. - Some benchmarks may not be interpretable to laypeople	- Proposed a framework for evaluating interactive LLMs like ChatGPT. - Evaluated ChatGPT's performance on 23 data sets covering 8 NLP tasks. - Found that ChatGPT outperforms LLMs with zero-shot learning on most tasks.
14	Yushi Bai et al.	- LMExamQA dataset construction- Evaluation metric design- Peer-examination pipeline	- - Potential bias during evaluation due to different model preferences and biases. Lack of robust evaluation capability among existing foundation models for large-scale peer examination.	- Language-Model-as-an-Examiner framework for benchmarking foundation models - Comprehensive evaluation with questions across multiple domains and follow-up questions - Combination of scoring and ranking measurements for reliable - results

15	Md Tahmid et al.	- Zero-shot evaluation of ChatGPT using benchmark datasets and tasks. - Leaderboard-based evaluation and task-based evaluation. Human intervention and automatic metrics used for evaluation.	- Unknown instruction-tuning datasets of OpenAI models - Numerical results may change as OpenAI trains new models - Lack of log-probability ranking- based evaluation - Limited details about evaluation approach in other - LLM papers	- Conducts a comprehensive evaluation of ChatGPT on bench- mark datasets. - Investigates effectiveness and limitations in various scenarios. - Studies language understanding and generation capability, commonsense reasoning, and open domain knowledge.
16	Percy Liang et al.	- Taxonomize scenarios and metrics of interest for language models. - Measure 7 metrics for each of 16 core scenarios. Conduct targeted evaluations based on 26 specific scenarios.	- Three categories of limitations: results, benchmark implementation, and benchmark design principles. - Lack of coverage of what is missing in - the implementation.	- Taxonomize scenarios and metrics for language models. - Multi-metric approach with 7 metrics for 16 core scenarios. - Large-scale evaluation of 30 language models - on 42 scenarios.
17	Potsawee Manakul et al.	- SelfCheckGPT with Prompt - SelfCheckGPT with BERTScore SelfCheckGPT with QA	- Study focused on passages about individuals, could be extended to other con- cepts. - Factuality con- sidered at sentence level, but sentences can have mixed information. - SelfCheckGPT with Prompt is computationally heavy, - needs improvement.	- Proposes SelfCheckGPT, a sampling-based approach for detecting hallucinated or factual responses generated by LLMs. - Shows that SelfCheckGPT is highly effective in hallucination detection and can outperform greybox methods.

18	Sewon Min et al.	- Methods include in-context learning, editing models, and semantic similarity score.	- FACTSCORE is not applicable for nuanced, open-ended, and debatable facts. - FACTSCORE may - not be suitable for nuanced human- written text with deception.	- Introducing FACTSCORE, a new evaluation method for factuality in long-form text generation. - Conducting extensive human evaluation of people biographies generated by commercial LMs. - Reporting the need for a fine-grained score in evaluating generation quality.
19	Chenyan Lyu et al.	Brainstorming interesting di- rections for MT using LLMs- Discussing privacy concerns in MT using LLMs- Proposing basic privacy-preserving methods to mitigate risks	- Designing intuitive and user-friendly interfaces for interactive MT. Incorporating user feedback into the translation process effectively.	Brainstormed interesting directions for MT using LLMs. - Explored stylized MT, interactive MT, and TM-based MT. - Proposed a new evaluation paradigm for translation quality using LLMs.
20	Cheng-wei Qin et al.	- Comparison of zero-shot learning performance of Chat-GPT and GPT-3.5 - Two-stage prompting method for zero-shot CoT Task instructions taken from or inspired by Brown et al. (2020)	- Excludes larger- scale datasets and more task categories due to cost limitations. - Reports best results for models not publicly available. Reports results based on best prompt found for public models.	- Systematic study of zero-shot learning capability of ChatGPT. - Evaluation on a large collection of NLP datasets covering 7 task categories. - Comparison of Chat- GPT performance with other models - Reporting results from recent work on zero-shot, fine-tuned, or few-shot fine-tuned. models
21	Cunx-iang Wang et al.	- Lexical matching - Neural evaluation Large language model (LLM)	- Data subject to frequent model up- dates, limiting reproducibility. - Unable to gather ample open-QA data based on GPT-4. - Limited labeling of dev set and train set due to resource limitations. - Risk of disseminating misinformation due to inaccuracies in gold standard answers.	- Introducing the task of Evaluating Open-QA Evaluation (QA-Eval). - Creating the dataset EVOUNA for evaluating AI-generated answers in Open-QA. - Investigating methods that show high correlation with human evaluations.

Ehsan Kamalloo et al. [19] concentrate on lexical matching and manual evaluation methods, emphasizing their efficacy for factoid information-seeking questions with short answers. However, their approach encounters challenges with more complex question forms and generative models, prompting the need for alternative evaluation methods such as regex matching, Zafaryab Rasool et al. [34] introduce the Cogtale dataset and evaluate LLM performance on various question types. Despite showcasing LLMs' strengths in single-choice and yes-no questions, they identify limitations in numerical extraction and diminished performance on certain question types. Ella Rabinovich et al. [32] contribute by creating a benchmark dataset with paraphrases and proposing a semantic consistency metric. Nevertheless, their approach relies on external knowledge, posing challenges when such knowledge is unavailable, Tal Schuster et al. [36] delve into semi-extractive multi-source QA (SEMQA) and hierarchical table regularization. While their contributions expand the scope of QA tasks, they acknowledge limitations related to model hallucinations and the confined nature of their datasets. The subsequent paper [39]introduces an LLM-based toolkit for TableQA, showcasing its potential with a comprehensive benchmark. However, they acknowledge the challenge of setting proper epochs in instruction tuning.

Zishan Guo et al. [15] address safety concerns related to LLMs, emphasizing the need for a comprehensive evaluation framework. Safety issues, potential misuse, and the absence of safeguards are highlighted, urging the community to address these critical concerns. Shahriar Golchin et al. [14] propose the Data Contamination Quiz for detecting and estimating data contamination in LLMs. Their approach, utilizing a four-option multiple-choice format, proves effective without requiring access to pre-training data, Ning Bian et al. [7] investigate the commonsense abilities of GPTs, revealing their struggles with certain types of knowledge. Despite achieving good accuracy in commonsense QA tasks, the authors identify areas where improvement is needed. Yejin Bang et al.[5] focus on evaluating ChatGPT's performance on various NLP tasks, demonstrating its superiority in many scenarios. However, concerns about limited sample sizes and potential safety issues are acknowledged, Yushi Bai et al. [4] contribute by introducing the LMEx-amQA dataset and a Language-Model-as-an-Examiner framework for benchmarking foundation models. They emphasize the need for a robust eval- uation capability and combine scoring and ranking measurements for reliable results. Md Tahmid et al. [22] conduct a zero-shot evaluation of ChatGPT, exploring its effectiveness and limitations in various scenarios. Their comprehensive evaluation includes benchmark datasets and investigates language understanding, generation capability, commonsense reasoning, and open domain knowledge.

Percy Liang et al. [23] taxonomize scenarios and metrics for language models, conducting a large-scale evaluation of 30 models on 42 scenarios. They identify limitations in bench- mark implementation and highlight the importance of addressing what is missing. Potsawee Manakul et al. [26] propose SelfCheckGPT for detecting hallucinated or factual responses, demonstrating its effectiveness. They discuss the computational heaviness of SelfCheckGPT and the need for improvement, Sewon Min et al. [27] introduce FACTSCORE, a new evaluation method for factuality in long-form text generation. Their extensive human evaluation of biographies generated by commercial LLMs reveals the need for fine-grained scores in evaluating generation quality. Chenyang Lyu et al. [25] brainstorm interesting directions for machine translation (MT) using LLMs and propose basic privacy-preserving methods. Their evaluation paradigm for translation quality emphasizes intuitive interfaces and user feedback incorporation, Chengwei Qin et al. [31] systematically study the zero-shot learning capability of ChatGPT, evaluating it on a large collection of NLP datasets. Their comparison with other models underscores the importance of reporting results from recent work on zero-shot, fine-tuned, or few-shot fine-tuned models. Cunxiang Wang et al. [38] introduce the task of Evaluating Open-QA Evaluation (QA-Eval) and the EVOUNA dataset for evaluating AI-generated answers in Open-QA. They explore

methods showing high correlation with human evaluations, although challenges in gathering ample open-QA data and potential misinformation dissemination are acknowledged.

7. FINDINGS

In this section will answer the research questions in below:

Answer 1, Architectural Symphony and Training's Tune: The choice of architecture and training objectives isn't just a technicality; it's a dance that dictates how LLMs perform on different QA tasks. Transformer-based models, with their masterful grasp of con- text, waltz across tasks like open-ended questioning and summarization. They effortlessly weave long threads of information, understanding subtle relationships. Meanwhile, recur- rent neural networks, though less graceful with distant connections, can tango expertly with sequential information, excelling in dialogue generation and sentiment analysis. Training objectives act as the choreographer, shaping the model's focus. Accuracy-driven training equips an LLM for the precision of closed-domain, factual questions, while reasoning-focused training molds it into a logician, adept at multi-hop retrieval and problem-solving. Comparing performances across architectures and objectives requires task-specific benchmarks, the equivalent of discerning judges, to truly appreciate the nuances of each dance.

Answer 2, Size, Time, and Data - The LLM Power Pack: Data, training time, and model size are the fuel that propels LLMs towards QA mastery. Imagine a vast reservoir of data – the bigger it is, the further an LLM can travel towards accuracy, though returns diminish as the reservoir overflows. Training time becomes the engine, allowing the LLM to refine its understanding with each turn of the crank. Longer durations enable deep exploration of complex relationships, but at the cost of increased fuel consumption. And finally, model size acts as the engine's capacity. Bigger models can hold more information and navigate intricate connections, but they guzzle resources and risk overfitting, akin to overloading an engine. Compared to traditional techniques, LLMs often zoom past in ac- curacy and reach, especially with a full tank. However, traditional methods, like nimble motorcycles, can still navigate specific terrains more efficiently due to their interpretability and fuel efficiency. Optimizing data usage, employing transfer learning, and model pruning techniques become the tuning dials, maximizing performance while conserving resources.

Answer 3, Unveiling the Intricacies: LLMs Confronting Complexity: When it comes to intricate questions, LLMs possess a unique lens. Their versatility allows them to tackle diverse question types, like puzzles with numerous pieces. Their contextual under- standing, like a magnifying glass, lets them sift through vast amounts of text, piecing together information to formulate comprehensive answers. They adapt readily, shifting strategies like a chameleon to fit new domains and tasks. However, weaknesses lurk in the shadows. Reasoning and inference, akin to solving intricate locks, can pose significant challenges. Factual accuracy can be a tightrope walk, with biases inherited from their training data potentially causing missteps. Finally, their reasoning can be shrouded in mystery, like a hidden code, making user interaction a delicate dance. While LLMs excel at deciphering complexities and crafting creative responses, traditional methods, like trusty compasses, can prove more reliable for navigating factual terrain and making logical deductions.

Answer 4, Bias Unveiled: Mitigating the LLM Shadow: Biases, like unwanted guests, can infiltrate LLM outputs, leading to unfair or discriminatory answers. Social biases, akin to warped mirrors, can reflect skewed perspectives related to gender, race, or ethnicity. Knowledge biases, like blind spots, can distort results due to unequal representation of topics or viewpoints in their

training data. To combat these unwanted guests, we need a multi-pronged approach. Data-driven methods, like filtering, augmentation, or reweighting, act as bouncers, weeding out bias from the data pool. Algorithmic approaches, akin to bias detectors, can be incorporated into training objectives or post-processing outputs to identify and neutralize bias. Finally, human oversight, like watchful hosts, remains crucial, especially for high-stakes applications, to identify and correct biased outputs. Measuring and identifying biases remains a challenge, requiring further research and development of robust evaluation methods, the equivalent of sophisticated scanners, to fully illuminate the LLM landscape.

6.1. Future Directions for Research

Large Language Models (LLMs) have revolutionized the field of natural language processing and question answering (QA), demonstrating remarkable capabilities in understanding and generating human-like text Brown et al. [9]. As we reflect on the current state of LLMs, it becomes evident that several exciting avenues for future research promise to further advance the capabilities and applications of these powerful models, and While pre-trained LLMs exhibit impressive generalization, fine-tuning for domain-specific QA remains an area ripe for exploration. Research efforts could focus on developing more effective fine-tuning strategies to enable LLMs to adapt seamlessly to specialized domains, ensuring more accurate and contextually relevant responses in fields such as healthcare, law, or finance Liu et al. [24]. The current landscape of LLMs excels in single-turn tasks but faces challenges in maintaining context and coherence across multiple turns in interactive and conversational QA scenarios. Future research should explore methods to improve interactive and conversational question-answering techniques for dynamic conversation tracking and context retention to enhance LLMs' abilities to engage in meaningful and coherent multi-turn dialogues Adiwar dana et al.[1], And building trust in QA systems is crucial for broader adoption. Also should interpret QA Models to prioritize the development of LLMs that not only provide accurate answers but also offer explanations for their responses. An explanatory approach would con- tribute to a deeper understanding of model reasoning and foster trust among users Jiang et al.[17], Also as LLMs are trained on vast amounts of data from the internet, the mitigating and ethical considerations is important. So, addressing biases and ethical concerns becomes paramount, And it should delve into techniques for debiasing LLMs and ensuring fair and unbiased answers, minimizing the perpetuation of societal biases present in training data Bolukbasi et al.[8]. Expanding LLMs to process and understand multimodal information presents a com- pelling research direction. Investigating methods for seamless integration of images, audio, or video into LLMs can open new possibilities for answering questions that involve information beyond textual content, therefore multimodal QA integration making QA systems more versatile Tan and Bansal et al [37]and the impressive zero-shot and few-shot learning abilities of LLMs are areas that can be further enhanced, At the end it could focus on improving the understanding of prompts and refining few-shot learning strategies, contributing to more effective and versatile QA systems Schick and Schutze. [35].

7. CONCLUSION

In conclusion, the future of QA with LLMs holds immense promise for advancements that will redefine natural language understanding. These research directions not only seek to improve the technical capabilities of LLMs but also address critical issues related to trans- parency, fairness, and user-centricity, contributing to the development of more robust and reliable QA systems, The rate of advancement of LLMs has been remarkable, showcasing significant progress across multiple tasks. However, despite ushering in a new era of artificial intelligence, our comprehension of this innovative form of intelligence remains relatively limited. It is crucial to establish the limitations of these LLMs' capabilities, comprehend their performance in different

domains, and investigate how to utilize their potential more effectively. This necessitates the implementation of a comprehensive benchmarking framework to guide the course of LLMs' development. This analysis systematically elaborates on the fun- damental abilities of LLMs, encompassing crucial aspects such as knowledge and reasoning. Moreover, we delve into the evaluation of alignment and safety, which includes ethical con- siderations, biases, toxicity, and truthfulness, in order to ensure the secure, trustworthy, and ethical application of LLMs. Concurrently, we explore the potential applications of LLMs in diverse fields, including biology, education, law, computer science, and finance. Most notably, we present a variety of widely-used benchmark evaluations to assist researchers, developers, and practitioners in comprehending and assessing the performance of LLMs. We anticipate that this analysis will encourage the development of evaluations for LLMs, providing clear guidance to steer the controlled progress of these models. This will enable LLMs to better serve the community and the world, ensuring that their applications in various domains are safe, dependable, and beneficial. With eager anticipation, we embrace the future challenges of LLMs' development and evaluation.[11]

ACKNOWLEDGEMENTS

I would like to thank Deanship of scientific research in King Saud University for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR).

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Razvan Hallak, Noah Fiedel, Romal Thoppilan, and Ilya Sutskever. Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977, 2020.

[2] Daman Arora, Himanshu Gaurav Singh, et al. Have llms advanced enough? a challenging problem solving benchmark for large language models. arXiv preprint arXiv:2305.15074, 2023.

[3] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language- model-as-an-examiner. arXiv preprint arXiv:2306.04181, 2023.

[4] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. Benchmarking foundation models with language-model-as-an-examiner. 2023.

[5] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. Centre for Artificial Intelligence Research (CAiRE), The Hong Kong University of Science and Technology, 2023.

[6] Yonggu Bang, Surya Cahyawijaya, Nayeon Lee, Weizhe Dai, Dawei Su, Brian Wilie, Hana Lovenia, Zeyu Ji, Tian Yu, Woojong Chung, Quoc Viet Do, Yanzhao Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023, 2023.

[7] Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. Journal of Artificial Intelligence Research, 2023.

[8] Tolga Bolukbasi et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. (Please provide the complete details of the journal or pub- lication), 2016.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and Dario Amodei. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

[10] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. 2023.

[11] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In Web, 2017.

[12] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. arXiv preprint arXiv:2304.00723, 2023.

[13] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. Instructeval: To- wards holistic evaluation of instruction-tuned large language models. arXiv preprint arXiv:2306.04757, 2023.

[14] Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. Journal of Computer Science, 2023.

[15] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey. Journal Name, 2023.

[16] Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. Evaluation of chatgpt on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. arXiv preprint arXiv:2306.04504, 2023.

[17] Ming Jiang et al. Explainable ai for question answering: A review. (Please provide the complete details of the journal or publication), 2021.

[18] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke S. Zettlemoyer. Spanbert: Im- proving pre-training by representing and predicting spans. 2020.

[19] Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. Journal Name, 2023. * Corresponding author: ekamalloo@uwaterloo.ca.

[20] Taichi Kojima and Shuai Shane Gu. Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 2022.

[21] Catherine Kosten, Philippe Cudr´e-Mauroux, and Kurt Stockinger. Spider4sparql: A complex benchmark for evaluating knowledge graph question answering sys- tems. 2023. Emails: {Catherine.Kosten@zhaw.ch, Philippe.Cudre-Mauroux@unifr.ch, Kurt.Stockinger@zhaw.ch}.

[22] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. 2023.

[23] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, and Kumar. Holistic evaluation of language models. 2023. Reviewed on OpenReview: https://openreview.net/ forum?id=iO4LZibEqW.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019.

[25] Chenyang Lyu, Jitao Xu, and Longyue Wang. New trends in machine translation using large language models: Case examples with chatgpt. 2023.

[26] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. 2023.

[27] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mo- hit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. 2023.

[28] Sewon Min, Percy Liang, and Caiming Xiong. Openqa: Open domain question answer- ing with large-scale knowledge bases. 2021.

[29] Linyong Nan, Ellen Zhang, Weijin Zou, Yilun Zhao, Wenfei Zhou, and Arman Co- han. On evaluating the integration of reasoning and action in llm agents with database question answering. 2023. Emails: {linyong.nan, ellen.zhang}@yale.edu.

[30] Dongqi Pu and Vera Demberg. Chatgpt vs human-authored text: Insights into control- lable text summarization and sentence style transfer. 2023.

[31] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? 2023.

[32] Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. Predicting question-answering performance of large language models through semantic consistency. Journal Name, 2023.

[33] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, 2018.

[34] Zafaryab Rasool, Scott Barnett, Stefanus Kurniawan, Sherwin Balugo, Rajesh Vasa, Courtney Chesser, and Alex Bahar-Fuchs. Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset. Journal Name, 2023.

[35] Timo Schick and Hinrich Schu¨tze. It's not just size that matters: Small language models are also few-shot learners. (Please provide the complete details of the journal or publication), 2020.

[36] Tal Schuster, Adam D. Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William W. Cohen, and Donald Metzler. Semqa: Semi-extractive multi-source question answering. Journal Name, 2023.

[37] Hui Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. 2019.

[38] Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. Evaluating open-qa eval- uation. 2023.

[39] Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Empower large language model to perform better on industrial domain-specific question answering. Journal Name, 2023.

[40] Wenting Zhao, Ye Liu, Tong Niu, Yao Wan, Philip S. Yu, Shafiq Joty, Yingbo Zhou, and Semih Yavuz. Divknowqa: Assessing the reasoning ability of llms via open-domain ques- tion answering over knowledge base and text. 2023. Emails: {wzhao41, psyu}@uic.edu, wanyao@hust.edu.cn, {yeliu, tniu, sjoty, yingbo.zhou, syavuz}@salesforce.com.

[41] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. 2023. Emails: {yczhuang, yueyu, kuanwang, haotian.sun, chaozhang}@gatech.edu.