



# Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family

Yiming Tan<sup>1,4</sup>, Dehai Min<sup>2,4</sup>, Yu Li<sup>2,4</sup>, Wenbo Li<sup>3</sup>, Nan Hu<sup>2,4</sup>,  
Yongrui Chen<sup>2,4</sup>, and Guilin Qi<sup>2,4(✉)</sup>

<sup>1</sup> School of Cyber Science and Engineering, Southeast University, Nanjing, China  
[tt\\_yymm@seu.edu.cn](mailto:tt_yymm@seu.edu.cn)

<sup>2</sup> School of Computer Science and Engineering, Southeast University, Nanjing, China  
[{zhishanq,yuli\\_11,nanhu,yrchen,gqi}@seu.edu.cn](mailto:{zhishanq,yuli_11,nanhu,yrchen,gqi}@seu.edu.cn)

<sup>3</sup> School of Computer Science and Technology, Anhui University, Hefei, China  
[wenboli@stu.ahu.edu.cn](mailto:wenboli@stu.ahu.edu.cn)

<sup>4</sup> Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University, Ministry of Education, Nanjing, China

**Abstract.** ChatGPT is a powerful large language model (LLM) that covers knowledge resources such as Wikipedia and supports natural language question answering using its own knowledge. Therefore, there is growing interest in exploring whether ChatGPT can replace traditional knowledge-based question answering (KBQA) models. Although there have been some works analyzing the question answering performance of ChatGPT, there is still a lack of large-scale, comprehensive testing of various types of complex questions to analyze the limitations of the model. In this paper, we present a framework that follows the black-box testing specifications of CheckList proposed by [38]. We evaluate ChatGPT and its family of LLMs on eight real-world KB-based complex question answering datasets, which include six English datasets and two multilingual datasets. The total number of test cases is approximately 190,000. In addition to the GPT family of LLMs, we also evaluate the well-known FLAN-T5 to identify commonalities between the GPT family and other LLMs. The dataset and code are available at <https://github.com/tan92hl/Complex-Question-Answering-Evaluation-of-GPT-family.git>.

**Keywords:** Large language model · Complex question answering · Knowledge base · ChatGPT · Evaluation · Black-box testing

---

Y. Tan and D. Min—Contribute equally to this work.

## 1 Introduction

Given its extensive coverage of knowledge from Wikipedia as training data and its impressive natural language understanding ability, ChatGPT has demonstrated powerful question-answering abilities by leveraging its own knowledge. Additionally, a study conducted by [30] suggests that language models can be considered as knowledge bases (KBs) to support downstream natural language processing (NLP) tasks. This has led to growing interest in exploring whether ChatGPT and related large language models (LLMs) can replace traditional Knowledge-Based Question Answering (KBQA) models.

There have been many evaluations of ChatGPT [2, 7, 16, 19, 26, 33, 46, 47, 52, 54], some of which include the testing of question answering tasks and have yielded interesting conclusions: for example, [26] showed that ChatGPT has lower stability than traditional KBQA models on a test set of 200 questions, and [2] found that ChatGPT is a “lazy reasoner” that suffers more with induction after analyzing 30 samples. However, due to the limited number of test cases, it is difficult to perform a comprehensive evaluation of ChatGPT’s performance on the KBQA task based on these findings. Moreover, the reliability of these findings still requires further testing for validation. We find that the difficulty in answer evaluation is the main reason why existing works have not conducted large-scale KBQA tests on ChatGPT, which outputs sentences or paragraphs that contain answers rather than an exact answer. Furthermore, due to the influence of the generated textual context, the answer sequence of ChatGPT may not necessarily correspond strictly to entity names in the knowledge base. Therefore, the traditional Exact Match (EM) metric cannot directly evaluate the output of ChatGPT for question-answering. Consequently, most of the works mentioned above rely on manual evaluation.

In this paper, we select the KB-based Complex Question Answering (KB-based CQA) task to comprehensively evaluate the ability of LLMs to answer complex questions based on their own knowledge. This task requires the model to use compositional reasoning to obtain the answer to the question, which includes multi-hop reasoning, attribute comparison, set operations, and other complex reasoning. We believe that evaluating ChatGPT’s performance in complex knowledge question answering using its own knowledge can help us understand whether existing LLMs have the potential to surpass traditional KBQA models or whether ChatGPT is already capable of replacing the current best KBQA models. Therefore, we collect test data from existing KB-based CQA datasets and establish an evaluation framework.

Our evaluation framework consists of two parts: 1) the feature-driven unified labeling method is established for the KBQA datasets involved in the testing; and 2) the evaluation of answers generated by LLMs. Inspired by the approach of using multiple scenario tags to evaluate language models in the HELM framework [21], we label each test question with unified answer-type, inference-type, and language-type tags. In the answer evaluation part, we first improve the Exact Match (EM) method so that it can be used to evaluate the accuracy of LLMs’ output. The main process of improved EM is to extract potential answer

phrases from the LLM output through constituent trees as the candidate answer pool, and then match them with the reference answer pool formed by annotated answers and aliases provided by wikidata. Next, we follow the CheckList testing specification [38] and set up three tests: the minimal functionality test (MFT), invariance test (INV) [40], and directional expectation test (DIR). Along with an overall evaluation, these tests assess the LLMs' capability, stability, and control when answering questions and performing specific reasoning operations.

Finally, we collect six English real-world KB-based CQA datasets and two multilingual real-world KB-based CQA datasets for our evaluation experiment, with a scale of approximately 190,000 questions, including approximately 12,000 multilingual questions covering 13 languages. In the experiment, we mainly compare the QA performance differences between the traditional the current state-of-the-art (SOTA) models and the GPT family models [4, 27, 28]. In addition, we also introduce the open-source LLM FLAN-T5 [9] model as a representative of the non-GPT family for comparison. Like ChatGPT, all the LLMs involved in the comparison in this paper use their own knowledge to answer questions and are considered unsupervised models.

Our key findings and insights are summarized as follows:

ChatGPT and the LLMs of GPT family outperform the best traditional models on some old datasets like WQSP and LC-quad2.0, but they still lag behind the current state-of-the-art on the latest released KBQA database such as KQAPro and GrailQA.

GPT family LLMs and the FLAN-T5 model tend to have similar tendencies in terms of strengths and weaknesses when answering different types of questions.

Using chain-of-thought prompts in CheckList testing enhances GPT LLMs' ability to answer specific questions but may negatively impact other question types, suggesting their potential and sensitivities for future task-specific applications.

## 2 Related Work

### 2.1 Large Language Models and Prompting

In recent years, LLMs and prompt learning have attracted considerable attention. Groundbreaking studies such as [4, 17, 30] revealed that LLMs, when given appropriate textual prompts, can perform a wide range of NLP tasks with zero-shot or few-shot learning without gradient updates. On the one hand, improved prompting can enable the information contained in the LLM to be more accurately applied to the target task, and early representative works include [34, 37]. The chain-of-thought (CoT) [48] method is a distinguished approach in effective prompt research. CoT enables LLMs to have a better understanding and think more when answering questions. On the other hand, much work has been done to improve the natural language understanding ability of LLMs, including Gopher [35] and PaLM [8], which aim to extend LLMs. Undoubtedly, ChatGPT has garnered significant attention as a prominent LLM due to its remarkable

natural language understanding abilities. It is trained on the GPT-3.5 series of models [11] using RLHF.

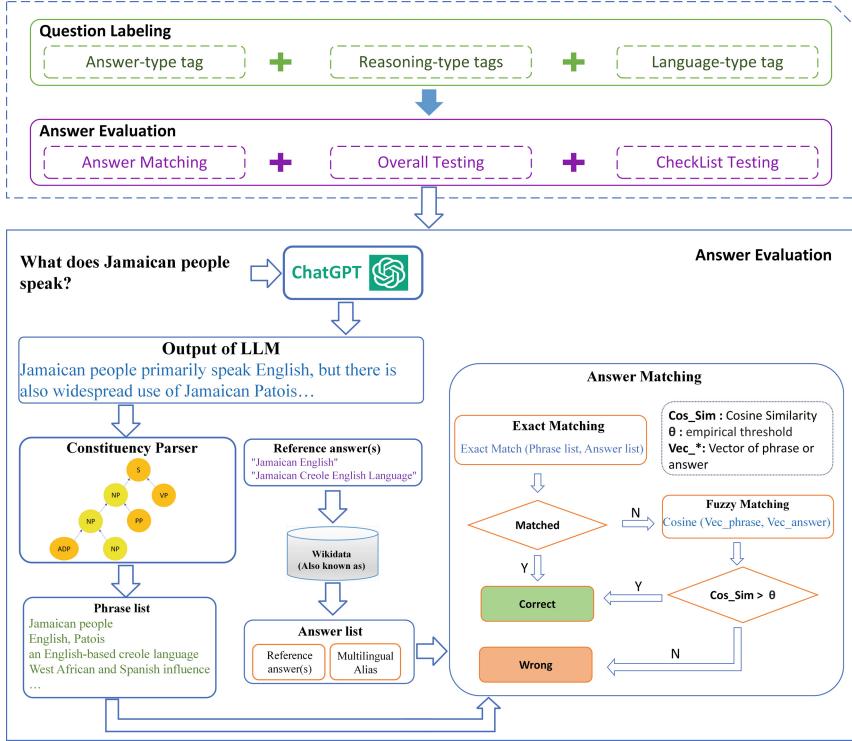
## 2.2 Evaluation of the Large Language Model

While LLMs have demonstrated outstanding natural language understanding and generation capabilities, it is still necessary to further research their strengths, limitations, and potential risks to fully understand their advantages. Recently, many works aimed at evaluating LLMs have been proposed [6], including general benchmarks like HELM [21], Bigbench [41], Promptbench [53], and MME [10]. These aim to categorize and summarize multiple existing tasks, providing a macro-level assessment of LLM performance and potential biases. Other studies focus on specific NLP tasks, such as summarization [2], question-answering [1, 2, 26], and machine translation [23]. In these existing works, the advantages of the general benchmark approaches lie in their fine-grained sample classification and high testing efficiency. However, these benchmarks are limited by the use of automated metrics, which restrict the diversity of testing objectives. On the other hand, evaluating task-specialized LLMs introduces more manually defined testing objectives, such as interpretability, determinism, robustness, and question understanding. Nevertheless, due to manual testing costs, these evaluations often rely on small samples (less than 10k) and coarsely categorized datasets.

In this paper, we combine the strengths of both benchmark studies and task-specific manual evaluations to test the GPT family LLMs. To achieve this, we adopt a strategy inspired by HELM [21], which uses multiple feature labels to describe and categorize task types, especially complex problem types. Additionally, we incorporate the manually predefined testing objectives from [26] and combine them with the CheckList natural language model’s black-box testing strategy. This comprehensive and diverse testing approach allows us to draw more comprehensive and valuable conclusions.

## 2.3 Black-Box Testing of the NLP Model

The prohibitive expense associated with training LLMs renders white-box testing an impractical approach. Consequently, the majority of assessment efforts presently concentrate on black-box evaluation approaches for LLMs. For example, the methods used by [3, 39] for evaluating robustness, the methods used by [49] for adversarial changes, and attention and interpretability within LLMs research conducted by [45]. The most comprehensive approach currently available is the CheckList approach proposed by [38], which categorizes evaluation targets into three parts: the minimum functionality test (MFT), invariance test (INV), and directional expectation test (DIR). The MFT examines a model’s basic functionality, INV examines whether the model can maintain functional correctness when non-answer-affecting information is added to the input, and DIR examines whether the model can output the expected result when the input is modified. In this work, we follow the idea of CheckList and use CoT prompting to generate test cases for DIR.



**Fig. 1.** Overview of proposed Evaluation Framework.

### 3 Evaluation Framework

As mentioned in Sect. 1, our KBQA evaluation framework consists of two parts. The first part aims to assign uniform feature labels to the questions in the datasets. The second part includes an improved Exact Match answer evaluation strategy and an extended CheckList test. Figure 1 illustrates the overall process of the framework. The detailed process is described in the following section.

#### 3.1 Feature-Driven Unified Question Labeling

We collect multiple existing KB-based CQA datasets for the evaluation. However, due to the different annotation rules used for features such as answer and reasoning type in each dataset, we need to establish a standardized and unified set of question feature tags for evaluating and analyzing question types.

Referring to the question tags provided by existing KBQA datasets [5, 22, 24, 51], we categorize the tags that describe the features of complex questions into three types, including answer type, reasoning type and language type. Table 1 lists the eight answer type tags and seven reasoning type tags we defined. Generally, a question contains one answer type tag, one language type tag and several

**Table 1.** The feature-driven question tags defined in this paper.

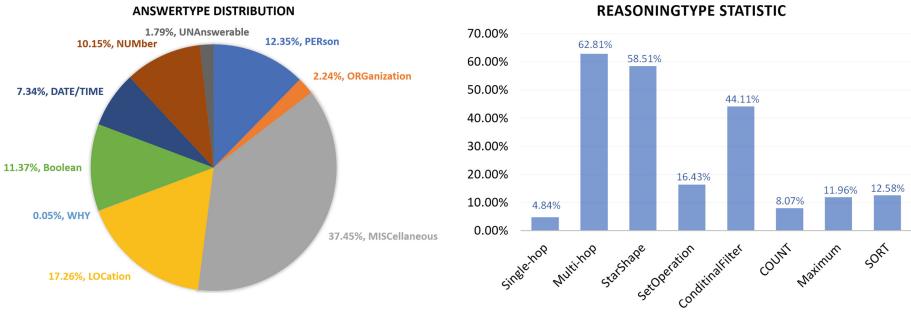
Answer type	Description
MISC	The answer to the question is the miscellaneous fact defined by the named entity recognition task
PER	The answer to the question is the name of a person
LOC	The answer to the question is a location
WHY	The answer explains the reasons for the facts mentioned in the question
DATE	The answer to the question is a date or time
NUM	The answer to the question is a number
Boolean	The answer to the question is yes or no
ORG	The answer to the question is the name of a organization
UNA	The input question is unable to answer.
Reasoning type	Description
SetOperation	The process of obtaining answers involves set operations
Filter	The answer is obtained through condition filtering
Counting	The process of obtaining an answer involves counting operations
Comparative	The answer needs to be obtained by comparing or sorting numerical values
Single-hop	Answering questions requires a single-hop Reasoning
Multi-hop	Answering questions requires multi-hop Reasoning
Star-shape	The reasoning graph corresponding to inputting question is star-shape

reasoning type tags. Figure 2 presents the label distribution of the data collected in this paper. For an input question, our labeling process is as follows: when the dataset provides question type tags, we simply match them to our feature tag list. When no tag is provided, we use an existing bert-base-NER model [18,44] to identify the type of answer, and use keywords in SPARQL to identify the type of inference.

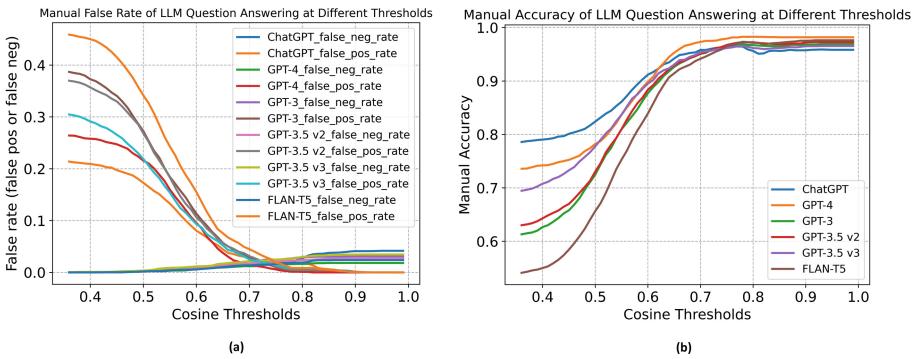
### 3.2 Answer Evaluation

The output of traditional KBQA models typically takes two forms: either a SPARQL query or a precise answer. The evaluation strategy for traditional KBQA models is based on exact match (EM), which involves comparing the model's output with a reference answer or to assess its accuracy. However, without adding additional prompts, LLMs generate text paragraphs containing answers, rather than precise answers. Furthermore, this answer may be a restatement of the reference answer.

**Extended Answer Matching.** To obtain evaluation results on KBQA outputs of LLMs resembling exact match, we propose an extended answer matching



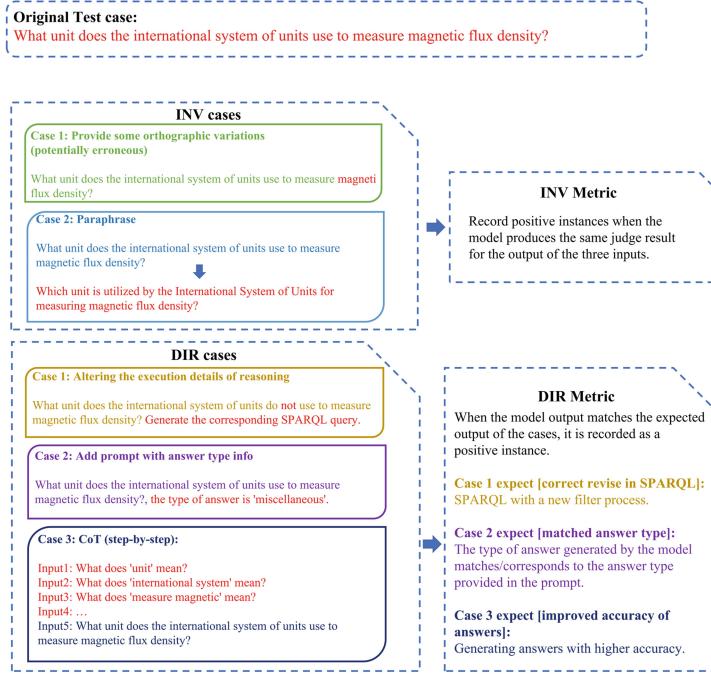
**Fig. 2.** The distribution of feature labels in the collect KB-based CQA datasets



**Fig. 3.** (a) The GPT family and T5 models show changing error rates on sampled questions as the threshold varies. (b) LLMs' QA accuracy (evaluated manually) on sampled questions varies with the threshold.

approach. This approach consists of three main parts: 1) Parsing LLMs' output using constituent trees [14] to extract NP or VP root node phrases as the candidate answer pool. 2) Expanding each reference answer using multilingual alias lists from Wikidata, including various names and aliases. 3) Using m-bert [18] to calculate the maximum Cosine similarity between reference and candidate answers for precise match evaluation, with a fuzzy matching strategy applied only to non-“NUM, DATE, Boolean” answer types.

**Threshold Selection and Sensitivity Analysis.** As shown in Fig. 3 (a), the analysis of various models reveals that using only EM evaluation for answers (threshold = 1) may result in 2.38%–4.17% (average 3.89%) false negative cases. To address this issue, we opt for establishing a fuzzy matching process based on cosine similarity to alleviate the problem. However, selecting an inadequate threshold may introduce additional false positive issues. Therefore, we followed the steps below to find an empirical threshold that minimizes the overall false rate (false pos + false neg) across all models: (1) We randomly sampled 3000 question samples from the test data of answer types involved in fuzzy matching



**Fig. 4.** Test cases design for INV and DIR.

and manually verified the correctness of the six LLM output answers shown in Fig. 3(a) (binary labels, correct/incorrect). (2) We calculate the minimum cosine similarity (the value is 0.38) between the gold answer and its aliases, and used it as the lower bound for finding the threshold. (3) We observed the changes in false rates for each model as the threshold increased from 0.38 to 1 and selected the threshold of 0.78 that minimized the average false rate across models. From the Fig. 3(a), it can be observed that the false rates of each model stabilize around this value. To evaluate the sensitivity of model performance to the threshold, as shown in Fig. 3(b), we compared the accuracy of each model on the test data as the threshold varied. The accuracy of each model tended to stabilize when the threshold was  $>0.7$ . Finally, we use 0.78 as the empirical threshold for further experiments. Sampling tests show that this threshold decreases the average false rate from 3.89% to 2.71%.

### 3.3 Checklist Testing

Following the idea of Checklist, we also evaluate ChatGPT and other LLMs with three distinct objectives: (1) to evaluate the ability of LLMs to handle each feature in KB-based CQA through the MFT; (2) to evaluate the robustness of LLMs' ability to handle various features in KB-based CQA scenarios through the INV; and (3) to evaluate whether the outputs of LLMs meet human expectations

for modified inputs through the DIR, the controllability. The specific INV and DIR procedures are presented as follows, and Fig. 4 presents the instances:

**Minimum Functionality Test.** In this work, we choose to examine the performance of LLMs in performing basic reasoning tasks by only including questions that involve a single type of reasoning operation. We then compare and analyze the performance differences of the models in answering questions that require performing a single reasoning operation versus those that require performing multiple reasoning operations.

**Invariance Test.** We design two methods to generate test cases for INV: the first method is to randomly introduce spelling errors into the original sentence, and the second method is to generate a question that is semantically equivalent (paraphrased) to the original sentence. Subsequently, we evaluate the invariance of the LLMs by checking the consistency of their correctness in the outputs generated from three inputs, i.e. the original test sentence, the version of the question with added spelling errors, and the paraphrased question.

**Directional Expectation Test.** In this study, we designed three modes for DIR test cases: (1) Replacing phrases related to reasoning operations in questions, to observe if LLMs' outputs correspond to our modifications. (2) Adding prompts with answer types after the original question text to check LLMs' ability to control the output answer type. (3) Using multi-round questioning inspired by CoT, where LLMs consider information related to key nouns before asking the original question, to observe the effectiveness and sensitivity of CoT prompts for different question types.

## 4 Experiments

### 4.1 Datasets

To highlight the complexity of the testing questions and the breadth of the testing dataset, after careful consideration, we selected six representative English monolingual KBQA datasets and two multilingual KBQA datasets for evaluation. These datasets include classic datasets such as WebQuestionSP [51], ComplexWebQuestions [43], GraphQ [42] and QALD-9 [24], as well as newly proposed datasets such as KQAPro [5], GrailQA [12] and MKQA [22]. Due to the limitations of the OpenAI API, we sampled some datasets, such as MKQA (sampled by answer type) and GrailQA (only using the test set). The collection size for each dataset and the scale we collected are summarized in Table 2.

### 4.2 Comparative Models

**State-of-the-Art Models for Each Dataset.** We introduce current SOTA models' report scores from the KBQA leaderboard [29] for each dataset as traditional KBQA models in this paper for comparison. This primarily reflects the comparison between LLMs and traditional KBQA models in terms of the overall results.

**Table 2.** The Statistical of collected KB-based CQA datasets, “Col. Size” represents the size of the dataset we collected in our experiments. “Size” denotes the original size of the dataset.

Datasets	Size	Col. Size	Lang
KQapro	117,970	106,173	EN
LC-quad2.0	26,975	26,975	EN
WQSP	4737	4,700	EN
CWQ	31,158	31,158	EN
GrailQA	64,331	6,763	EN
GraphQ	4,776	4,776	EN
QALD-9	6,045	6,045	Mul
MKQA	260,000	6,144	Mul
Total Collected		194,782	

**Large-Language Models of the GPT Family.** ChatGPT is a landmark model in the GPT family, and we believe that comparing it to its predecessors and subsequent versions is very valuable. By doing so, we can observe and analyze the technical increments of the GPT family at each stage and the benefits they bring. In this paper, we compare the GPT family models, which include GPT-3, GPT-3.5 v2, GPT-3.5 v3, ChatGPT (Their names on OpenAI’s Model Index document are: text-davinci-001, text-davinci-002, text-davinci-003, gpt-3.5-turbo-0301) and the newest addition, GPT-4 [27].

**Large-Language Model Not Belongs to GPT Family.** The LLM we have chosen is the famous FLAN-T5 (Text-to-Text Transfer Transformer 11B, [7]), which does not belong to the GPT family. Considering its multilingual question-answering ability and open-source nature, we have chosen it to participate in the comparison in this paper. FLAN-T5 is an encoder-decoder transformer language model that is trained on a filtered variant of CommonCrawl (C4) [36]. The release date and model size for this model are also based on [36].

### 4.3 Overall Results

The overall results are presented in Table 3. First, ChatGPT outperforms the current SOTA traditional models on three of the eight test sets, and the subsequently released GPT-4 surpasses on four test sets. By comparing the performance of GPT-4 and SOTA models, we can see that as LLMs represented by the GPT family, their zero-shot ability is constantly approaching and even surpassing traditional deep learning and knowledge representation models.

Second, comparing models in the GPT family, the newer models perform better than the previous ones, as expected. Interestingly, the performance improvement of the new GPT models is relatively consistent across all datasets, as shown in Fig. 5(a), where the line shapes of all GPT models are almost identical. This means that each generation of GPT models retains some commonalities. Based

**Table 3.** Overall results of the evaluation. We compare the exact match of ChatGPT with current SOTA traditional KBQA models (fine-tuned (FT) and zero-shot (ZS)), GPT family LLMs, and Non-GPT LLM. In GraphQ, QALD-9 and LC-quad2, the evaluation metric used is F1, while other datasets use Accuracy (Exact match).

Datasets	KQapro	LC-quad2	WQSP	CWQ	GrailQA	GraphQ	QALD-9	MKQA
	Acc	F1	Acc	Acc	Acc	F1	F1	Acc
SOTA(FT)	<b>93.85</b> [29]	33.10 [31]	73.10 [15]	<b>72.20</b> [15]	<b>76.31</b> <sup>‡</sup>	31.8 [13]	<b>67.82</b> [32]	46.00 [22]
SOTA(ZS)	94.20 [25]	-	62.98 [50]	-	-	-	-	-
FLAN-T5	37.27	30.14	59.87	46.69	29.02	32.27	30.17	20.17
GPT-3	38.28	33.04	67.68	51.77	27.58	38.32	38.54	26.97
GPT-3.5v2	38.01	33.77	72.34	53.96	30.50	40.85	44.96	30.14
GPT-3.5v3	40.35	39.04	79.60	57.54	35.43	47.95	46.19	39.05
ChatGPT	47.93	42.76	83.70	64.02	46.77	53.10	45.71	44.30
GPT-4	57.20	<b>54.95</b>	<b>90.45</b>	71.00	51.40	<b>63.20</b>	57.20	<b>59.20</b>

<sup>‡</sup> <https://dki-lab.github.io/GrailQA/>.

**Table 4.** Comparison of LLMs on multilingual test sets.

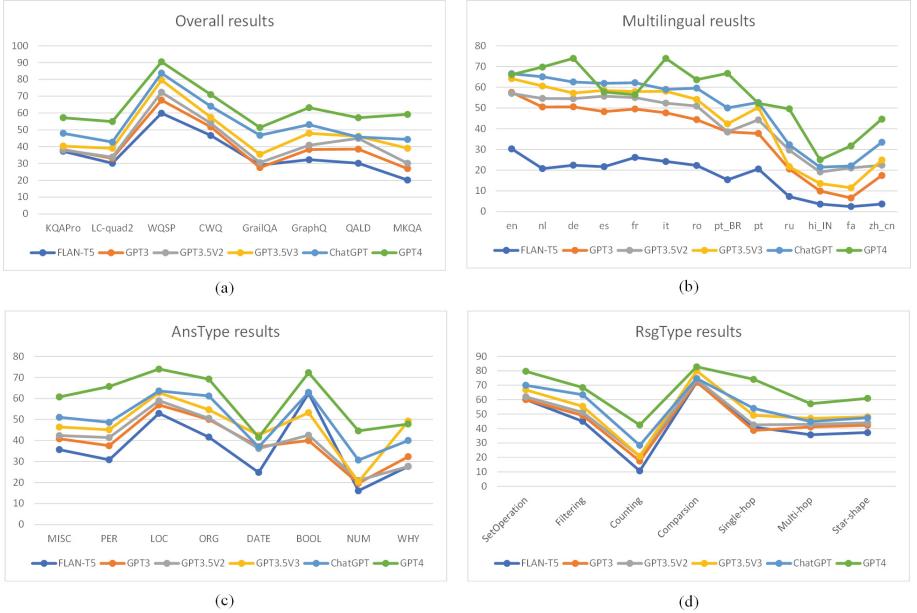
Languages	FLAN-T5	GPT-3	GPT-3.5v2	GPT-3.5v3	ChatGPT	GPT-4
en	30.29	57.53	56.99	64.16	<b>66.49</b>	66.09
nl	20.75	50.47	54.58	60.56	65.05	<b>69.72</b>
de	22.40	50.54	54.48	57.17	62.54	<b>73.91</b>
es	21.68	48.22	55.70	58.50	<b>61.87</b>	57.69
fr	26.16	49.46	55.02	57.89	<b>62.19</b>	62.00
it	24.19	47.67	52.33	58.06	58.96	<b>73.91</b>
ro	22.28	44.38	50.94	54.12	59.55	<b>63.41</b>
pt_br	15.38	38.46	38.46	42.31	50.00	<b>66.67</b>
pt	20.58	37.70	44.26	50.27	<b>52.64</b>	52.25
ru	7.29	20.58	29.69	21.68	32.24	<b>49.58</b>
hi_in	3.61	9.93	19.13	13.54	21.48	<b>25.00</b>
fa	2.45	6.59	21.09	11.49	22.03	<b>31.71</b>
zh_cn	3.65	17.45	22.40	24.87	33.46	<b>44.62</b>

on the known cases, these commonalities may come from the transformer-based encoding. We will discuss in detail the impact they have in Sect. 4.5. In addition, we can observe that the newer versions of the GPT model show increasingly significant improvements compared to the previous generations.

Third, as shown in Fig. 5(a), although FLAN-T5’s overall performance is weaker than that of the GPT family, its line shape is quite similar to that of the GPT family. This further supports our inference that the transformer-based architecture leads to commonalities in the abilities of current LLMs.

#### 4.4 Multilingual KBQA Results

Based on the results from MKQA and QALD-9, we further present the performance of LLMs on multilingual QA in Table 4. Despite the overall trend showing



**Fig. 5.** (a) is the line chart based on Table 3, showing the EM scores of each model on different datasets. (b) corresponds to Table 5, with lines representing the EM scores of each model in different languages. (c) and (d) correspond to Table 4, reflecting the trend of EM scores of each model on different types of questions.

improvement in the model's ability to answer questions in different languages as the GPT family continues to iterate, we observe that GPT-4 has not surpassed ChatGPT in the four languages. This suggests that the evolution of GPT's multilingual capabilities may be starting to slow down. Figure 5(b) shows the line chart of the EM scores of the models in each language. We can find that the shapes of the lines for GPT-3 and ChatGPT are very similar, while there is a significant change in the shape of the line for GPT-4. We believe that the main reason for this change is due to the introduction of multimodal data in GPT-4, which plays a positive role in mapping between some languages.

#### 4.5 Feature Tags Based Results

The results in Table 5 show the performance of ChatGPT and other LLMs when answering different types of questions. As such, traditional models are not compared in this section. Overall, models from the GPT family are better at answering questions with boolean (yes/no) answers, questions about organizations and locations, as well as those involving set operations and numerical comparisons. However, they do not perform well when answering questions that require precise dates or involve numerical calculations. From the performance of models in the GPT family and Flan-T5, it can be found that Flan-T5 performs worse

**Table 5.** Exact Match comparison based on Answer Types (AnsType) and Reasoning Types (RsgType)

MF	FLAN-T5	GPT-3	GPT-3.5v2	GPT-3.5v3	ChatGPT	GPT-4
AnsType						
MISC	35.67	40.79	42.35	46.42	51.02	<b>60.73</b>
PER	30.84	37.53	41.36	45.10	48.65	<b>65.71</b>
LOC	52.91	56.92	58.93	62.71	63.55	<b>73.98</b>
ORG	41.62	50.01	50.58	54.62	61.18	<b>69.20</b>
DATE	24.81	37.07	36.15	<b>42.54</b>	36.92	41.57
Boolean	62.43	39.96	42.56	53.23	62.92	<b>72.28</b>
NUM	16.08	19.66	21.01	20.31	30.70	<b>44.59</b>
WHY	27.69	32.31	27.69	<b>49.23</b>	40.00	47.83
UNA	-	-	-	-	-	-
RsgType						
SetOperation	60.11	60.12	62.03	66.86	70.00	<b>79.70</b>
Filtering	45.01	49.06	51.24	55.43	63.40	<b>68.40</b>
Counting	10.68	17.56	20.83	20.83	28.41	<b>42.50</b>
Comparison	72.13	72.44	74.00	80.00	74.74	<b>82.79</b>
Single-hop	41.00	38.72	42.54	49.22	54.00	<b>74.14</b>
Multi-hop	35.68	41.09	42.98	47.06	44.88	<b>57.20</b>
Star-shape	37.23	42.28	43.96	48.17	47.43	<b>60.91</b>

in all cases except for questions with boolean answer types. This is consistent with the conclusion of [21]: The performance of knowledge-intensive tasks is closely related to the size of the model. For comparisons within the GPT family of models, following the iteration process of the GPT model summarized in [11], we also observe some positive effects of certain technical introductions on the model, including: (1) [11] point out that GPT-3.5 v3 has better in-context learning abilities, while ChatGPT sacrifices these abilities in order to model dialogue history. This may explain why GPT-3.5 v3 performs better in answering multi-hop and star-shaped questions that require distinguishing entity mentions through context. (2) ChatGPT’s dialogue learning helps it better answer short questions (Single-hop). (3) The GPT-3.5 v2, obtained through language model and code training followed by supervised instruction tuning, but its overall capabilities do not appear to have significantly improved compared to GPT-3. The possible reason could be that alignment harms performance, and the alignment tax offsets the increase in zero-shot ability obtained through training [21, 28]. (4) One possible reason why the successor models outperform GPT3.5 V2 in most aspects is that the complex reasoning ability acquired through training on code, which did not manifest prominently in GPT3.5 V2, but were unlocked after the introduction of instruction tuning with RLHF [9, 28].

Figures 5(c) and (d) respectively show line chart of the EM scores formed by each model in answering questions of different answer and reasoning types. The two Figures are consistent with what we observed in the Overall results, that is, various models of the GPT family and FLAN-T5 have similar line shapes. In addition, we also find that the performance of the new GPT models has improved significantly in some specific types of questions, such as Boolean-type (ChatGPT, GPT-4) and WHY-type (GPT-3.5 v3). However, in some other types of questions, there is no significant improvement for the multi-generation models of GPT, such as Num-type and Counting-type. This indicates that there is still a significant room for improvement for LLMs, and the iteration is far from over. Another interesting finding is that FLAN-T5 performs similarly to ChatGPT in answering boolean questions, but performs worse than the GPT family in other types of answers. Due to the difference in their training data, we cannot accurately determine in the current evaluation whether the reason for this situation is the difference in training data or whether certain training strategies used by the GPT family have a negative impact on specific types of questions.

**Table 6.** MFT results of ChatGPT

	SetOperation	Filtering	Counting	Comparison	Single-hop	Multi-hop	Star-shape
Single Reasoning	60.22	51.39	24.16	31.48	44.07	<b>48.27</b>	<b>50.75</b>
Multiple Reasoning	<b>70.00</b>	<b>63.40</b>	<b>28.41</b>	<b>74.74</b>	<b>54.00</b>	44.88	47.43

## 4.6 CheckList Results

**MFT Results.** In the MFT tests, we only evaluate questions that contain a single type of reasoning or multiple reasoning labels of the same type (such as SetOperation+Comparison and SetOperation+Filtering). Based on the results of MFT, we compared the performance of ChatGPT in answering single and multiple reasoning questions. Table 6 shows the following findings. (1) Except for multi-hop and star type questions, ChatGPT performs better in executing multiple reasoning than in performing single reasoning in answering questions involving other types of reasoning operations. (2) ChatGPT is not good at answering counting questions despite the improvements generated by multiple reasoning.

**INV Results.** Table 7 presents the stability of LLMs from the GPT family across three runs on three different test cases. As a reference, [26] noted that the stability of traditional KBQA models is 100. The results in Table 7 are reported using the following symbols: ‘CCC’ indicates that all answers to the three inquiries are correct, while ‘WWW’ indicates that none of the three

**Table 7.** INV results of GPT family

LLM	CCC	CCW	CWC	CWW	WCC	WCW	WWC	WWW	Stability Rate
GPT-3	434	64	59	52	42	43	73	666	76.76
GPT-3.5 v2	495	44	65	42	43	30	58	656	80.30
GPT-3.5 v3	604	46	43	49	34	35	49	583	82.83
ChatGPT	588	49	72	68	52	27	32	545	79.06
GPT-4	798	0	0	65	54	0	0	516	<b>91.70</b>

**Table 8.** DIR results for RsgType, the score represents the percentage of expected output produced by the LLMs.

	SetOperation	Filtering	Counting	Comparison	Overall
GPT-3.5 v3	45%	75%	65%	<b>65%</b>	62.5%
ChatGPT	<b>75%</b>	85%	<b>70%</b>	<b>65%</b>	<b>73.75%</b>
GPT-4	65%	<b>90%</b>	<b>70%</b>	60%	71.25%

inquiries received correct answers or the model did not return any useful answers. Only when the correctness of the three queries is consistent, the model's performance on the problem is considered stable. As shown in Tables 7, the overall stability of the GPT models has improved from GPT-3 to GPT-4, and GPT-4 has reached a stability rate of 91.70, which is very close to that of traditional KBQA models. The stability rate of ChatGPT is slightly lower than that of GPT-3.5, and we infer that this is due to the fact that the ChatGPT model focuses more on conversation training, resulting in higher instability (randomness) in the output.

**DIR Results.** As mentioned in Fig. 4, we designed three DIR modes to examine the controllability of LLMs from the GPT family. In the first mode, we manually observe whether the SPARQL statements output by the model contain the expected keywords, and calculate the failure rate of the model's expected reasoning operations. Since GPT-3.5 v2 and its earlier versions did not undergo code learning, it is difficult for them to generate correct SPARQL queries. Therefore, in this test, we compare the GPT-3.5 v3, ChatGPT, and GPT-4. As shown in Table 8, the scores of around 73% indicate that even the latest GPT model still has a high degree of randomness in performing reasoning operations, which will affect its applicable scenarios.

In the second mode, we provide prompts to the model's input indicating the answer type and observe the change in the EM score. In Table 9, red values indicate that adding prompts increases the EM score, while blue values indicate negative effects. For most models, prompts have a relatively stable positive effect on Boolean and NUM type questions, while the answers to MISC type questions are mostly negatively affected. In addition, in new models such as ChatGPT and GPT-4, the effect of answer type prompts is much worse than in GPT-3.5 and earlier models. This suggests that different models have different internal

**Table 9.** DIR results for AnsType prompting

	MISC	PER	LOC	ORG	DATE	Boolean	NUM	WHY
GPT-3	+1.43	0	+5.71	+4.29	+4.29	+15.71	+17.14	0
GPT-3.5 v2	-4.28	+2.85	+7.14	+14.28	+2.86	-8.57	+14.28	+12.13
GPT-3.5 v3	-12.86	+10.00	+18.57	-7.14	+4.71	+17.14	+22.85	+9.09
ChatGPT	+6.78	-3.64	-1.72	-5.35	-8.58	+4.28	+7.15	-3.03
GPT-4	-4.29	-2.86	+11.43	+5.71	0	+7.14	+4.29	-6.06

**Table 10.** DIR results for CoT prompting

	MISC	PER	LOC	ORG	DATE	Boolean	NUM	WHY
GPT-3	-1.40	-2.00	-2.67	+2.73	-3.77	+3.36	+35.66	+6.06
GPT-3.5 v2	-0.35	-5.33	+1.78	-3.64	+0.76	-5.04	+32.95	0
GPT-3.5 v3	0	-2.00	-1.33	-1.82	-1.51	-2.10	+34.12	0
ChatGPT	-1.75	-4.66	+0.89	-3.63	-1.50	+3.36	+30.62	+6.06
GPT-4	-3.00	+11.11	+2.22	+3.3	-2.71	0	+20.00	+2.62
	SetOperation	Filtering	Counting	Comparison	Multi-hop	Star-shape		
GPT-3	+10.79	+10.43	+35.66	+1.35	-1.60	-1.69		
GPT-3.5 v2	+4.86	+5.46	+38.54	-2.26	-1.18	-0.85		
GPT-3.5 v3	+6.34	+8.18	+38.99	-1.13	-1.61	-1.26		
ChatGPT	+7.82	+9.47	+35.78	+0.45	-1.47	-1.41		
GPT-4	+2.05	+0.93	+11.11	-1.88	+2.82	+2.68		

knowledge and understanding of the same input text, and the effectiveness and helpfulness of the same prompt vary among different models. More powerful models are more sensitive to the content of prompts because of their powerful natural language understanding ability. It may be difficult to design simple and universally effective prompts that work well across all models.

In the third mode, we guide the model step by step through a naive CoT-guided process to first provide the crucial information required to answer the question, and then answer the original question. Table 10 shows the difference in EM scores of the GPT model’s answers before and after using CoT-guided process for each type of questions. We can observe that positive impact brought by CoT to GPT-4 is greater than that of other models, and the improvement of CoT on the model’s ability to answer NUM-type questions is significant and stable. In terms of reasoning types, CoT improves the ability of all models in set operations, conditional filtering, and counting, but it doesn’t help much with multi-hop and star-shape questions. Specifically, the most significant improvement introduced by CoT for the GPT family of models is a score increase of over 30.00 for answer types that are numerical (NUM). This result strongly supports the importance of CoT for using LLMs to solve numerical-related questions [20].

## 5 Conclusion

In this paper, we extensively tested the ability of ChatGPT and other LLMs to answer questions on KB-based CQA datasets using their own knowledge. The experimental results showed that the question-answering performance and reliability of the GPT model have been continuously improving with version iterations, approaching that of traditional models. CheckList testing showed that current LLMs still have a lot of room for improvement in some reasoning abilities, and CoT-inspired prompts can improve the performance of the original model on certain types of questions. Consequently, this evaluation serves as a valuable reference for future research in the relevant community. In future work, we need to further expand on the following two points: Firstly, conduct tests in various domains to validate which conclusions obtained from open-domain KBQA are universal and which are domain-specific. Secondly, perform tests on various types of models. With ongoing LLM research, besides the GPT family, many new large-scale open-source models have been proposed. It requires further exploration and summarization to determine if they possess better mechanisms for self-knowledge organization or stronger capabilities to accept human prompts and find answers.

**Acknowledgments.** This work is supported by the Natural Science Foundation of China (Grant No. U21A20488). We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

## References

1. Bai, Y., et al.: Benchmarking foundation models with language-model-as-an-examiner. arXiv preprint [arXiv:2306.04181](https://arxiv.org/abs/2306.04181) (2023)
2. Bang, Y., et al.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv e-prints, arXiv-2302 (2023)
3. Belinkov, Y., Glass, J.: Analysis methods in neural language processing: a survey. Trans. Assoc. Comput. Linguist. **7**, 49–72 (2019)
4. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
5. Cao, S., et al.: KQA Pro: a dataset with explicit compositional programs for complex question answering over knowledge base. In: Proceedings ACL Conference, pp. 6101–6119 (2022)
6. Chang, Y., et al.: A survey on evaluation of large language models. arXiv preprint [arXiv:2307.03109](https://arxiv.org/abs/2307.03109) (2023)
7. Chen, X., et al.: How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. arXiv e-prints, arXiv-2303 (2023)
8. Chowdhery, A., et al.: PaLM: scaling language modeling with pathways. arXiv e-prints, arXiv-2204 (2022)
9. Chung, H.W., et al.: Scaling instruction-finetuned language models. arXiv preprint [arXiv:2210.11416](https://arxiv.org/abs/2210.11416) (2022)
10. Fu, C., et al.: MME: a comprehensive evaluation benchmark for multimodal large language models. arXiv preprint [arXiv:2306.13394](https://arxiv.org/abs/2306.13394) (2023)
11. Fu, Y., Peng, H., Khot, T.: How does GPT obtain its ability? Tracing emergent abilities of language models to their sources. Yao Fu’s Notion (2022)

12. Gu, Y., et al.: Beyond IID: three levels of generalization for question answering on knowledge bases. In: Proceedings WWW Conference, pp. 3477–3488 (2021)
13. Gu, Y., Su, Y.: ArcaneQA: dynamic program induction and contextualized encoding for knowledge base question answering. In: Proceedings COLING Conference, pp. 1718–1731 (2022)
14. He, H., Choi, J.D.: The stem cell hypothesis: dilemma behind multi-task learning with transformer encoders. In: Proceedings EMNLP Conference, pp. 5555–5577 (2021)
15. Hu, X., Wu, X., Shu, Y., Qu, Y.: Logical form generation via multi-task learning for complex question answering over knowledge bases. In: Proceedings ICCL Conference, pp. 1687–1696 (2022)
16. Huang, F., Kwak, H., An, J.: Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. arXiv e-prints, arXiv-2302 (2023)
17. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Trans. Assoc. Comput. Linguist. **8**, 423–438 (2020)
18. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings NAACL-HLT Conference, pp. 4171–4186 (2019)
19. Kocoń, J., et al.: ChatGPT: jack of all trades, master of none. arXiv e-prints, arXiv-2302 (2023)
20. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. arXiv preprint [arXiv:2205.11916](#) (2022)
21. Liang, P., et al.: Holistic evaluation of language models. arXiv e-prints, arXiv-2211 (2022)
22. Longpre, S., Lu, Y., Daiber, J.: MKQA: a linguistically diverse benchmark for multilingual open domain question answering. Trans. Assoc. Comput. Linguist. **9**, 1389–1406 (2021)
23. Lyu, C., Xu, J., Wang, L.: New trends in machine translation using large language models: case examples with ChatGPT. arXiv preprint [arXiv:2305.01181](#) (2023)
24. Ngomo, N.: 9th challenge on question answering over linked data (QALD-9). Language **7**(1), 58–64 (2018)
25. Nie, L., et al.: GraphQ IR: unifying the semantic parsing of graph query languages with one intermediate representation. In: Proceedings EMNLP Conference, pp. 5848–5865 (2022)
26. Omar, R., Mangukiya, O., Kalnis, P., Mansour, E.: ChatGPT versus traditional question answering for knowledge graphs: current status and future directions towards knowledge graph chatbots. arXiv e-prints, arXiv-2302 (2023)
27. OpenAI: GPT-4 technical report (2023)
28. Ouyang, L., et al.: Training language models to follow instructions with human feedback. arXiv e-prints, arXiv-2203 (2022)
29. Perevalov, A., Yan, X., Kovriguina, L., Jiang, L., Both, A., Usbeck, R.: Knowledge graph question answering leaderboard: a community resource to prevent a replication crisis. In: Proceedings LREC Conference, pp. 2998–3007 (2022)
30. Petroni, F., et al.: Language models as knowledge bases? In: Proceedings IJCAI Conference, pp. 2463–2473 (2019)
31. Pramanik, S., Alabi, J., Saha Roy, R., Weikum, G.: UNIQORN: unified question answering over RDF knowledge graphs and natural language text. arXiv e-prints, arXiv-2108 (2021)

32. Purkayastha, S., Dana, S., Garg, D., Khandelwal, D., Bhargav, G.S.: A deep neural approach to KGQA via SPARQL silhouette generation. In: Proceedings IJCNN Conference, pp. 1–8. IEEE (2022)
33. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is ChatGPT a general-purpose natural language processing task solver? arXiv e-prints, arXiv-2302 (2023)
34. Qin, G., Eisner, J.: Learning how to ask: querying LMS with mixtures of soft prompts. In: Proceedings NAACL-HLT Conference (2021)
35. Rae, J.W., et al.: Scaling language models: methods, analysis & insights from training gopher. arXiv e-prints, arXiv-2112 (2021)
36. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**(1), 5485–5551 (2020)
37. Reynolds, L., McDonell, K.: Prompt programming for large language models: beyond the few-shot paradigm. In: Proceedings CHI EA Conference, pp. 1–7 (2021)
38. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: behavioral testing of NLP models with checklist. In: Proceedings ACL Conference, pp. 4902–4912 (2020)
39. Rychalska, B., Basaj, D., Gosiewska, A., Biecek, P.: Models in the wild: on corruption robustness of neural NLP systems. In: Proceedings ICONIP Conference, pp. 235–247 (2019)
40. Segura, S., Fraser, G., Sanchez, A.B., Ruiz-Cortés, A.: A survey on metamorphic testing. *IEEE Trans. Software Eng.* **42**(9), 805–824 (2016)
41. Srivastava, A., et al.: Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv preprint [arXiv:2206.04615](https://arxiv.org/abs/2206.04615) (2022)
42. Su, Y., et al.: On generating characteristic-rich question sets for QA evaluation. In: Proceedings EMNLP Conference, pp. 562–572 (2016)
43. Talmor, A., Berant, J.: The web as a knowledge-base for answering complex questions. In: Proceedings ACL Conference, pp. 641–651 (2018)
44. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings NAACL-HLT Conference, pp. 142–147 (2003)
45. Wang, A., et al.: SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In: Proceedings NeurIPS Conference, pp. 3266–3280 (2019)
46. Wang, J., Liang, Y., Meng, F., Li, Z., Qu, J., Zhou, J.: Cross-lingual summarization via chatgpt. arXiv e-prints, arXiv-2302 (2023)
47. Wang, S., Scells, H., Koopman, B., Zuccon, G.: Can ChatGPT write a good Boolean query for systematic review literature search? arXiv e-prints, arXiv-2302 (2023)
48. Wei, J., et al.: Chain of thought prompting elicits reasoning in large language models. arXiv preprint [arXiv:2201.11903](https://arxiv.org/abs/2201.11903) (2022)
49. Wu, T., Ribeiro, M.T., Heer, J., Weld, D.S.: Errudite: scalable, reproducible, and testable error analysis. In: Proceedings ACL Conference, pp. 747–763 (2019)
50. Ye, X., Yavuz, S., Hashimoto, K., Zhou, Y., Xiong, C.: RNG-KBQA: generation augmented iterative ranking for knowledge base question answering. In: Proceedings ACL Conference, pp. 6032–6043 (2022)
51. Yih, W.T., Richardson, M., Meek, C., Chang, M.W., Suh, J.: The value of semantic parse labeling for knowledge base question answering. In: Proceedings ACL Conference, pp. 201–206 (2016)
52. Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D.: Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. arXiv e-prints, arXiv-2302 (2023)

53. Zhu, K., et al.: PromptBench: towards evaluating the robustness of large language models on adversarial prompts. arXiv preprint [arXiv:2306.04528](https://arxiv.org/abs/2306.04528) (2023)
54. Zhuo, T.Y., Huang, Y., Chen, C., Xing, Z.: Exploring AI ethics of ChatGPT: a diagnostic analysis. arXiv e-prints, arXiv-2301 (2023)