ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots

Reham Omar[1] , **Omij Mangukiya**[1] , **Panos Kalnis**[2] and **Essam Mansour**[1]
[1]Concordia University, Canada, [2]KAUST, Saudi Arabia
{Fname}.{Lname}@concordia.ca, panos.kalnis@kaust.edu.sa

Abstract

Conversational AI and Question-Answering systems (QASs) for knowledge graphs (KGs) are both emerging research areas: they empower users with natural language interfaces for extracting information easily and effectively. Conversational AI simulates conversations with humans; however, it is limited by the data captured in the training datasets. In contrast, QASs retrieve the most recent information from a KG by understanding and translating the natural language question into a formal query, supported by the database engine.

In this paper, we present a comprehensive study of the characteristics of the existing alternatives, towards combining both worlds into novel KG chatbots. Our framework compares two representative conversational models, ChatGPT and Galactica, against KGQAn, the current state-of-the-art QAS. We conduct a thorough evaluation using four real KGs across various application domains, to identify the current limitations of each category of systems. Based on our findings, we propose open research opportunities to empower QASs with chatbot capabilities for KGs. All benchmarks and all raw results are available[1] for further analysis.

1 Introduction

Both conversational AI and question-answering systems (*QAS*), aim to develop effective natural language interfaces for accessing useful information. Conversational AI models simulate conversations with humans to achieve a specific task, or participate in general discussions [Yan, 2018] [Young *et al.*, 2022] [Zhong *et al.*, 2021]. Their answers are formatted mainly as human-like text with explanations to build discussions. QASs focus mainly on retrieving information from KGs by translating a question into a structured query, that fetches the recent information relevant to the question from an existing KG [Omar *et al.*, 2023]. In QASs, an answer has a structured format, e.g., CSV- or Excel-like format, with no explanations. Combining the capabilities of conversational AI and question-answering systems for KGs, has the potential to lead the development of advanced chatbot systems. We envision KG chatbots that will enable companies to: (*i*) meet the diverse needs of end users; (*ii*) maximize the usability of their KGs; and (*iii*) maintain the KGs with dynamic context, based on the interactions with users.

There is a sharp growth in the use of conversational AI models, such as ChatGPT [2]. The model is trained on a large collection of datasets from diverse data sources, to generate human-like answers. Several KGs are captured from some of these datasets. This paper answers the question: *how does a language model, such as ChatGPT, compare against traditional QASs for KGs?* Our survey also identifies the main capabilities that should be combined into a chatbot for KGs. This new generation of KG chatbot systems will be able to benefit from the structured information in a KG, to provide more accurate answers based on a specific domain, such as academics, finance, or health care.

Future KG chatbot systems will need a QAS mechanism to update the response generated by language models with the recent information retrieved from a given KG. Thus, the user will get responses reflecting the most recent information related to their domain-specific questions. The state-of-the-art QAS for KG is KGQAn [Omar *et al.*, 2023], a universal QAS that can answer questions from arbitrary KGs of various application domains. In KGQAn, question-answering is formalized as a three-fold problem: question understanding, linking, and filtering of the final answers. For question understanding, KGQAn utilizes Seq2Seq pre-trained language models [Sutskever *et al.*, 2014] trained independently of KGs, such as BART [Lewis *et al.*, 2020] and GPT-3 [Brown *et al.*, 2020]. For linking and filtering, KGQAn uses a just-in-time approach based on word embedding models, such as FastText [Bojanowski *et al.*, 2017].

This paper proposes a comparative framework to systematically review the main capabilities of conversational AI language models versus KG QASs, and their styles in answering questions. We used our framework to comprehensively evaluate the state-of-the-art language models, such as ChatGPT and Galactica [3], versus the state-of-the-art QAS. Our evaluation uses four real KGs from various application domains. We identify the current limitations of

[1] https://gitfront.io/r/user-2881158/pCcUsY8d2145/lmkg/

[2] https://openai.com/blog/chatgpt/

[3] https://galactica.org

each category. We also propose open research challenges and opportunities for developing a KG chatbot that incorporates the merits of conversational AI models and QASs.

2 Background: Language Models and QASs

Language models, such as BERT [Devlin *et al.*, 2019], T5 [Raffel *et al.*, 2020], and GPT-3 [Brown *et al.*, 2020], solve the question-answering task using text generation. Some models need to be re-trained for specific tasks. Given the input, the pre-trained model will directly produce the output. Other models have to be fine-tuned using prompting, e.g., the model gets a description of the required task and a few examples; then it can directly be used for generating the required output. Based on the training datasets, language models can be classified into: (*i*) general models, such as PaLM [Chowdhery *et al.*, 2022], OPT [Zhang *et al.*, 2022] and GPT-NeoX-20B [Black *et al.*, 2022]; and (*ii*) specific domains models, such as Galactica [Taylor *et al.*, 2022], SciBERT [Beltagy *et al.*, 2019] and BioMegatron [Shin *et al.*, 2020].

More advanced conversational AI models and dialogue agents have benefited from the recent advancement in language models to create chatbots that can answer questions while engaging with users in conversations. Examples of these chatbots are ChatGPT, a recent chatbot introduced by OpenAI, and LaMDA [Thoppilan *et al.*, 2022], a family of transformer-based [Vaswani *et al.*, 2017] language models for dialogue applications. Moreover, Sparrow [Glaese *et al.*, 2022] is a dialogue agent trained using reinforcement learning with human feedback, and webGPT [Nakano *et al.*, 2021] is a fine-tuned GPT-3 model that answers questions by allowing the language model to search the web. Meena [Adiwardana *et al.*, 2020] is a chatbot trained to respond in a human-like way by replying using sensible responses.

ChatGPT is a chatbot based on language models for answering questions, asking for clarification, creating dialogues with the user, and dealing with follow-up questions. It can apply reasoning to correct its answer based on users' feedback. Moreover, ChatGPT can write, debug, and optimize code. It is trained using reinforcement learning with human feedback similar to the training process of InstructGPT [Ouyang *et al.*, 2022], but using a combination of (*i*) a new dialogue dataset, and (*ii*) InstructGPT dataset after transforming its format to a dialogue format.

Galactica is a general-purpose scientific language model trained on a large corpus of scientific data for multiple tasks, such as predicting citations, reasoning, question answering and predicting protein annotations. It is based on transformer models [Vaswani *et al.*, 2017] that consist of only a decoder with some modifications. It was trained on massive datasets from different open-access scientific sources, such as papers and filtered common crawl. Its training datasets also included some general knowledge, such as Wikipedia.

Traditional QASs for KGs divide the question-answering process into three phases: question understanding, linking, and execution. In the question understanding step, the systems extract key entities and relations from the given question. The entities and relations are organized in an intermediate representation of the question, usually in the form of a graph. In the linking step, QASs map the extracted key entities and relations to vertices and predicates from the target KG. The last step is organizing the retrieved vertices and predicates to create a SPARQL query, execute it and perform an optional filtration step, where the most relevant answers are returned to the user.

EDGQA [Hu *et al.*, 2021] uses Stanford core NLP parser [CoreNLP, 2022] to predict the constituency parsing tree of the question. Then it applies human-curated heuristic rules to transform the tree into a root-acyclic graph, called an entity description graph (EDG). In the linking step, the nodes of the EDG are linked to vertices in the target KG using different linking methods, such as Falcon [Sakor *et al.*, 2019]. These methods depend on building indices in a pre-processing phase, then using them to retrieve the required output. EDGQA also performs another pre-processing phase to extract all entity types from the KG for filtration. Finally, the semantically equivalent SPARQL query to the question Q is created using the linked vertices, predicates, and entity type to the different phrases in Q.

KGQAn [Omar *et al.*, 2023] is the current state-of-the-art system for question answering on KGs. Question understanding is formalized as a triple-patterns generation model, trained using Seq2Seq pre-trained models, such as BART [Lewis *et al.*, 2020] and GPT-3. The triple patterns are converted to a graph structure called Phrase Graph Pattern (PGP). KGQAn performs just-in-time linking based on built-in indices in the KG engines. It also uses word embedding models, such as FastText [Bojanowski *et al.*, 2017], to assess the semantic affinity between different phrases in the question and vertices and predicates in the KG. KGQAn does not depend on any pre-processing index, or models trained on the KG, to perform the linking. Thus, KGQAn can answer questions on an arbitrary KG as an on-demand service.

3 Comparative Framework

This section introduces our comparative framework towards chatbots for KGs. It also defines a unified assessment strategy to compare conversational AI language models and QA systems, e.g., deciding the correctness of the generated answers.

3.1 Criteria Towards KG Chatbots

There is a need for a comparative framework that evaluates the differences between conversational AI language models and QASs for the question answering (*QA*) task on KGs. We defined fundamental criteria, such as the answer's correctness, robustness, consistency, explainability and advanced understanding of different questions, plus more KG-oriented criteria, such as incorporating recent information and generality across domains.

Correctness: KGs are constructed by capturing information and semantics from different data sources for a specific domain. The QA task on KGs focuses more on extracting accurate answers that can be used as input data for subsequent tasks. Thus, the correctness of the provided answers is one of the essential criteria in our framework. We measure the correctness of the answers by comparing them to those included in a benchmark's ground truth. Our framework utilizes the micro-F1 score as the primary metric for correct answers. QASs, such as EDGQA and KGQAN, output the answers in a format that can be directly used by the evaluation script provided by QALD-9 [4]. This script automatically calculates the micro-F1 score of the systems and can be directly employed for the other datasets.

Language models, such as ChatGPT and Galactica, generate answers formatted differently than the answers in the KG benchmarks, as discussed more in Section 3.2. Thus, calculating the micro-F1 score for these models is performed manually. For example, the process may include comparing the provided answer to the labels of the KG vertices of the corresponding answers in the ground truth. To ensure a fair comparison with QA systems, we use the following definitions of the metrics, that are found in the QALD-9 script:

$$C = S \cap G \tag{1}$$

$$m_{prec} = \frac{\Sigma_{i=1}^{N} |C_{qi}|}{\Sigma_{i=1}^{N} |S_{qi}|} \tag{2}$$

$$m_{recall} = \frac{\Sigma_{i=1}^{N} |C_{qi}|}{\Sigma_{i=1}^{N} |G_{qi}|} \tag{3}$$

$$microF1 = \frac{2 \cdot m_{prec} \cdot m_{recall}}{m_{prec} + m_{recall}} \tag{4}$$

where S_q is the set of answers generated by the system for question q, G_q is the set of ground truth answers provided by the benchmark for question q, C_q is the set of correct answers generated by the system for question q, N is the number of questions in the benchmark, m_{prec} is the micro-precision, and m_{recall} is the micro-recall.

Determinism: KGs contain structured information that is queried using query languages, such as SPARQL, or Cypher. Thus, answering questions against a specific KG is expected to be deterministic, that is, for a given question, the generated answer will always be the same, assuming the same context. In contrast, language models encapsulate information seen in the training datasets, and there is a degree of non-deterministic behaviour in the generated answers. We consider this factor by running each question three times and calculating correctness based on the best answers received.

Robustness is the ability of KG chatbots to tolerate erroneous input, such as a question with spelling, or grammar mistakes; these are common due to the free-form input from humans. Natural language models are trained to generate the correct answers, despite the possible mistakes. In construct, in formal query languages, such as SPARQL, any simple mismatch in the vertices, or predicates will lead to wrong answers. Therefore, robustness is a challenging measure for QASs. Our framework evaluates robustness by injecting spelling, or grammar mistakes in a subset of the benchmarks' questions and evaluating the correctness of the answers.

Explainability: We define explainability as the ability to provide explanations, or additional details related to the question, which can benefit the user. The explanation increases the confidence of the users about the answer, as it explains the reasoning or the process followed to provide the answer. Explainability is common in language models. QASs, on the other hand, do not provide explanations, as their answers are extracted using SPARQL queries. In our framework, this factor is evaluated manually by observing the outputs of the language models for all benchmark questions.

Question understanding is the ability to understand a given question, regardless of the correctness of the answer. This criterion is essential, as users ask questions of different degrees of complexity, such as aggregate, temporal, or multiple-intentions questions. In QASs, question understanding is a separate phase followed by linking to entities in the KG. To evaluate the question understanding ability of such systems, we observe the output of the first phase in the pipeline. In contrast, language models encapsulate question understanding within the wider output generation process. Therefore, for a language model, we analyze the generated answer to decide whether the model understood the question.

Incorporating recent information: KGs are frequently updated by adding, deleting, or changing facts. For example, Wikipedia is updated at a rate of two edits per second[5], which results in hundreds of thousands of edits per day in the corresponding Wikidata KG. Users typically expect the most up-to-date answer. This criterion depends on the adopted process used to train the models, or the type indices used by the QASs. We consider this criterion in the evaluation.

Generality across different domains is one of the desirable criteria. We define it as the ability to support the question answering task on KGs of various domains without re-training on the target KG. Real KGs are of large scale. Thus, the training process for a specific KG will consume large amounts of time and computing resources. To evaluate generality, we use four real KGs of two domains and benchmarks, including human-curated questions of different complexities and styles.

3.2 Fairness and Manual Assessment

For the question answering task, language models and QASs may use different versions of the same knowledge, such as unstructured text versus structured data (graph). It is challenging to evaluate these models and systems based on the same criteria, due to the lack of benchmarks and a unified method for calculating the correctness of answers. To overcome these challenges, we performed manual evaluation while considering multiple factors to ensure the fairness of the assessment.

For correctness, we need to evaluate the outputs of the QASs and language models against the ground truth answers, which are either strings, or vertices from the KG. For QASs, this is straight-forward. In contrast, language models produce mainly human-like text. We compare the answer generated by language models to the label of the vertex from the KG. For example, for the question: *Give me all movies with Tom*

[4] https://github.com/ag-sc/QALD/blob/master/6/scripts/evaluation.rb

[5] https://en.wikipedia.org/wiki/Wikipedia:Statistics

Cruise, the output of the language models and the label of the ground truth is the titles of the movies, so they can be easily compared. However, for the question: *Who founded Intel*, language models produce the answer in the form of a paragraph mentioning the founders with some context related to the company. We manually search for the expected answer within that text and compare it to the ground truth. Each benchmark deals with a specific version of KGs. Hence, when evaluating language models, we only consider the answers in the period covered by the KG. For example, a particular benchmark uses a KG based on information until 2016 only. Some models, such as ChatGPT, encapsulate information until 2021. We exclude any answers not included in the KG, to assess the precision of the model fairly.

4 Experimental Evaluation

4.1 Compared Models and Systems

In our survey, we evaluate two traditional QASs for KG: KGQAn and EDGQA[6], against two language models: Galactica[7] and ChatGPT. KGQAn and EDGQA are the state-of-the-art systems for question answering on KGs. Galactica has five models with a different number of parameters. We used the standard model with 6.7B parameters. ChatGPT is a recent and very popular language model. We considered three variations of ChatGPT: (*i*) *Default*: we perform three runs per question using the "regenerate response" button; (*ii*) *Follow-up*: we ask a follow-up question that requests the entire list of answers instead of the examples provided, i.e., we first ask the question and, after retrieving the answer, we input *continue* to get a more detailed list; and (*iii*) *Excel*: we ask ChatGPT to return its answers in a structured format, like an excel sheet. At the start of each thread we write a predefined message and then, for each question, we append to it the prefix *List out all*. The message is as follows:

> You only form the answers to my questions as an excel table. Do not write explanations. I will ask you questions, fill out a sheet with the answer, and decide on rows and columns based on the answer. Give also index to each row incrementally in the sheet.

We refer to these three variations as ChatGPT, ChatGPT-Follow up and ChatGPT-Excel. In our experiments, answers are classified into three categories: (*i*) *Correct*, where ChatGPT provides correct answers with respect to the golden truth; (*ii*) *Wrong*, where ChatGPT provides answers that do not match the golden truth; and (*iii*) *No answer*, where ChatGPT is unable to produce any answer for the given question. The main reason for the failure is either ChatGPT was not trained on the given information, or it does not understand the entity we are asking about.

4.2 Benchmarks

Our evaluation utilizes four different benchmarks: two on general-fact KGs, namely DBpedia [Auer *et al.*, 2007] and YAGO[8] [Tanon *et al.*, 2020]; and another two benchmarks on KGs related to academia, namely DBLP[9] and the Microsoft Academic Graph (*MAG*)[10]. The general-fact KGs have facts about entities representing places and persons. DBLP and MAG contain facts about scientific publications, citations, authors and institutions. Each benchmark includes a set of English questions against one of these four KGs. Each question is annotated with the equivalent SPARQL query and the answers to the question. These answers are retrieved by executing the given SPARQL query. The benchmarks consist of different categories of questions, such as boolean questions whose answer is $True$ or $False$, listing questions where the answer should contain an exhaustive list of the coresponding KG entities, and WH-questions (e.g., WH-at, WH-o, etc.) that ask about people, places, languages, etc.

We use QALD-9 [Usbeck *et al.*, 2018], the most challenging and widely used benchmark to evaluate QASs. QALD-9 uses the DBpedia KG, and consists of 150 questions for testing. The questions of QALD-9 are human-generated. The other benchmarks, namely YAGO, DBLP, and MAG, are recently introduced by [Omar *et al.*, 2023]. Each of these benchmarks contains 100 questions for testing, which are also human-generated.The questions against YAGO are similar to the ones of QALD-9, that is, questions about people and places. Both DBLP and MAG benchmarks have questions related to citations and authors.

4.3 Performance Evaluation

We compare the two models and systems using the four benchmarks. Table 1 summarizes the precision, recall and micro F1 score for each competitor in each benchmark. KGQAn achieve comparable results on the general KGs (QALD-9 and YAGO) and the academic KGs (DBLP and MAG). ChatGPT performs significantly better on the general KGs compared to its performance on the academic KGs. ChatGPT does not add any irrelevant answers. Thus, ChatGPT is consistently achieving better precision than KGQAn on QALD-9 and YAGO. However, ChatGPT struggles in recall as it does not, by default, fully answer questions with a long list of answers. KGQAn utilized models trained independently of the target KG. Thus, KGQAn achieves better precision and recall for academic graphs than EDGQA, which is trained mainly on QALD-9. QALD-9 and YAGO use the general KGs. Hence, EDGQA achieves comparable results on both.

To benefit from the full potential of ChatGPT, we tested using the Excel and Follow-up variations of ChatGPT to retrieve more answers to questions with a long list of answers. ChatGPT-Excel achieved better results as we ask explicitly to *List out all* answers. As shown in Table 1, the ChatGPT- Excel improved upon the default ChatGPT for all benchmarks. It even surpassed KGQAn in YAGO benchmark. In general, ChatGPT-Follow-up outperformed ChatGPT-Excel, as we enforced further the model to provide the full answers. This shows that to get the full potential of ChatGPT, you

[6] https://github.com/HXX97/EDGQA/

[7] https://github.com/paperswithcode/galai

[8] https://yago-knowledge.org/downloads/yago-4

[9] https://dblp.org/rdf/release/dblp-2022-06-01.nt.gz

[10] https://zenodo.org/record/4617285#.YrNszNLMJhH

Table 1: Results of the four benchmarks. A QA system, such as KGQAn, is designed to provide full answers with high precision across KGs of different domains. Language models do not provide correct answers for domain-specific questions, where the training datasets did not cover.

	QALD-9			YAGO			DBLP			MAG		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
EDGQA	31.30	**40.30**	32.00	41.90	40.80	41.40	8.00	8.00	8.00	4.00	4.00	4.00
KGQAn	49.81	39.39	**43.99**	48.48	**65.22**	55.62	**57.86**	**52.02**	**54.78**	**55.41**	**45.61**	**50.04**
Galactica	14.19	2.31	3.97	9.92	3.48	5.15	1.50	0.42	0.66	18.02	0.02	0.04
ChatGPT	55.85	10.61	17.83	76.10	36.30	49.15	48.28	0.74	1.46	28.13	0.04	0.07
ChatGPT-Excel	52.08	13.65	21.63	79.95	43.25	56.14	45.76	2.85	5.37	32.67	0.08	0.16
ChatGPT-Follow up	**60.94**	14.44	23.34	**87.185**	52.29	**65.56**	51.52	3.59	6.72	33.14	0.06	0.11

Table 2: Performance of ChatGPT across three runs on QALD-9, where C means that the question was answered correctly, W means the question was not answered correctly and NA means that ChatGPT failed to return any useful answer.

Run 1	**Run 2**	**Run 3**	**Count**
C	C	C	83
C	C	W	1
C	W	C	1
C	W	W	3
NA	C	C	20
NA	C	W	2
NA	NA	C	1
NA	W	C	2
W	C	C	1
W	C	W	2
W	W	C	1
NA	NA	NA	4
NA	NA	W	1
NA	W	NA	1
NA	W	W	12
W	W	W	15

Table 3: Performance of ChatGPT across three runs on MAG, where C means that the question was answered correctly, W means the question was not answered correctly and NA means that ChatGPT failed to return any useful answer.

Run 1	**Run 2**	**Run 3**	**Count**
C	C	C	15
C	W	C	1
NA	NA	C	1
W	C	C	1
W	C	W	1
W	W	C	1
NA	NA	NA	53
NA	NA	W	1
NA	W	W	2
W	W	NA	1
W	W	W	23

need to engage in a conversation with it. Similar to ChatGPT, Galactica did not perform well on list questions, significantly affecting its recall and the overall F1 score. The performance of language models and systems like EDGQA dropped significantly in answering questions against unseen information.

4.4 Determinism

For the same question, KGQAn always produces the same SPARQL query and answers, regardless of the number of runs attempted. To evaluate ChatGPT, we run each question three times. In some questions, there is a slight change in the explanation of the answer, but the answer is the same. In other questions, the output changed utterly. In some cases, ChatGPT in the first run does not provide any answer but returns a correct answer in the second run. In other cases, It does not give any answer in the first run, but in the next run, it can produce an answer, but it is a wrong answer. Tables 2 and 3 show the outputs of ChatGPT in the three runs for both QALD-9 and MAG benchmarks. For lack of space, we removed the results of the other two benchmarks, which are similar to Tables 2 and 3. For QALD-9, ChatGPT is deterministic, producing the same answer 68%. For YAGO, it is deterministic in 85% of the cases. For DBLP and MAG, it is deterministic for 94% and 91% of the cases, respectively. Overall, ChatGPT achieves a high percentage of determinism across the benchmarks, while KGQAn is more deterministic than the language model.

4.5 Chatbot-oriented Criteria

Robustness: We chose five random questions that ChatGPT can answer. For each question, we modified it by introducing different spelling and grammatical mistakes. We produced five versions for each question and ran ChatGPT on them. It understood and solved all the modified questions except one. QASs, including KGQAn and EDGQA, cannot achieve robustness as spelling mistakes lead to a mismatch between extracted entities from a question and the corresponding vertices in a KG.

Explainability: KGQAn provides the user with the answer to the question. There is no way for the user to understand the process that led to receiving this answer. In contrast, ChatGPT provides an answer with a reasonable explanation in a human-like text. Sometimes, extra helpful information related to the answer is provided to the user.

Question Understanding Our comprehensive evaluation included 450 questions against different domains and KGs. We evaluated the correctness by manually comparing the generated answers to the ground truth. Through this process, we recognized that ChatGPT has a perfect understanding of questions, even for those it fails to answer. We observed

Table 4: Number of questions answered correctly per category, where SF is single fact questions, SFT are single facts questions with type, MF are multi-facts questions, and B are boolean questions. ChatGPT has better performance than KGQAN in understanding complex questions, such as SFT and MF. However, KGQAN performs better than ChatGPT across different domains, i.e., general and academic KGs.

Systems	**QALD-9 (150 Q)**				**YAGO (100 Q)**				**DBLP (100 Q)**				**MAG (100 Q)**			
Category	**SF**	**SFT**	**MF**	**B**	**SF**	**SFT**	**MF**	**B**	**SF**	**SFT**	**MF**	**B**	**SF**	**SFT**	**MF**	**B**
# of ques.	**81**	**28**	**37**	**4**	**87**	**6**	**6**	**1**	**85**	**11**	**4**	**0**	**75**	**7**	**16**	**2**
KGQAN	46	7	9	0	**61**	**5**	2	0	**49**	**4**	**1**	_	**40**	**2**	**9**	**2**
ChatGPT-Follow up	**68**	**19**	**27**	**3**	52	4	2	**1**	11	1	1	_	16	0	4	0

multiple answers to questions it was unable to answer, e.g. for question: *How many papers did Du Linna publish?* that ChatGPT fail to respond, its reply mentions:

> It is possible that you can find information about an individual's publications by searching online or by contacting their place of employment or academic institution.

This answer is provided for many questions to help the user find the required information. We conclude that ChatGPT understands the question, but its training datasets do not include the necessary information. QASs, including KGQAN and EDGQA, lack these capabilities.

We also analyzed both ChatGPT-Follow up and KGQAN in answering questions of different linguistic complexity. Table 4 summarizes our results and shows the number of correctly answered questions for each category per benchmark. This subset covers all questions in the benchmarks, where SF are single fact questions, SFT are single fact questions with type, MF are multi facts questions, and B are boolean questions. For ChatGPT, we consider the question correctly answered if it produces the correct answers in any of the three runs. Although ChatGPT has good performance concerning the number of questions answered in the general knowledge benchmarks, this is not reflected in the F1 score. This is due to the low recall, as shown in Table 1.

4.6 KG-oriented Criteria

Unlike ChatGPT, KGQAN understands most of the questions of different types across the different domains and maintains comparable performance in precision, recall and F1 score. This proves the generality of KGQAN across different domains and users, who may express questions of variant complexity. As shown in Table 4, ChatGPT has a good performance on the general benchmarks, QALD-9 and YAGO, solving more than 50% of the questions. However, it cannot work as well in the academic KGs solving a maximum of 20% of the benchmark correctly. We can conclude that the data sources that ChatGPT is trained on do not contain enough academic information from DBLP and MAG KGs.

KGQAN can deal with new and updated KGs instantaneously. The user needs to provide KGQAN with the URL of the SPARQL endpoint of the new graph. The KGQAN models are independent of the KGs. KGQAN uses these models to convert a question into a SPARQL query, which will be executed against the KG and return an answer incorporating the recent information. ChatGPT is a trained model, which answers questions based on the seen data in training. If the data source is updated, ChatGPT will require further training to answer questions about the updated information.

Table 5: Comparison between ChatGPT and KGQAN according to the metrics of the proposed comparative framework. KG Chatbots needs a combination of language models' abilities, such as robustness and explainability, with QAS's abilities, such as generality across domains and providing recent information.

Metric	**ChatGPT**	**KGQAN**
Correctness	low	high
Robustness	Yes	No
Determinism	high	high
Explainability	Yes	No
Question Understanding	advanced	Basic
Incorporate recent information	No	Yes
Generality / different domains	low	high

4.7 Discussion

We summarize our comprehensive evaluation in Table 5 for ChatGPT and KGQAN based on our comparative framework. Both can answer a high percentage of questions on the general knowledge graph. ChatGPT has a lower recall for questions having a long list of answers. KGQAN has good performance accross different KGs, while ChatGPT could not answer most of the questions against DBLP and MAG. Both have Deterministic behaviour in answering questions. ChatGPT is slightly less deterministic than KGQAN, which incorporates recent information in the KG without needing further training. ChatGPT is trained on information from different data sources. Thus, it does not generalize to unseen domain information. In contrast, KGQAN can only retrieve answers from the target KG. ChatGPT has an outstanding ability in terms of explainability and robustness. There is a room for improvement for QASs, such as KGQAN, in terms of explainability and robustness and question understanding.

5 KG Chatbots: Open Research Challenges

Advanced KG Chatbots could be developed by incorporating capabilities of language models, such as ChatGPT, in question understanding, robustness, and explainability, with the outstanding capabilities of a QA system, such as KGQAN, in incorporating recent information and generality across different domains. This section discusses several research challenges in advancing QA systems with these capabilities.

5.1 Dialogue Support for KGs

The traditional QAS pipeline consists of three phases: (*i*) Question Understanding, (*ii*) Linking, (*iii*) Execution and Filtration. We need to modify question understanding and linking steps to handle the dialogue. For question understanding, most systems generate an intermediate representation of the question, e.g., a graph structure. Systems must enrich and maintain this intermediate representation as the dialogue progresses. It will need to specify if the new question in the conversation refers to an existing component or if we need to add a new node or edge to be up to date with the status of the conversation. It will also need to decide whether an answer to a previous question needs to replace a node in the graph. In this case, systems need to specify a new node representing the answer the user wants from the KG. QASs will have to link the newly added nodes and edges to vertices and predicates from the target KG.

The existing benchmarks for the QA task on KGs are not designed to evaluate dialogues. Instead, these benchmarks contain individual questions. The initial step of creating a chatbot for KGs is engaging in a conversation with the user. To achieve this, we need a new generation of dialogue benchmarks. The answers throughout the dialogue should be from the KG. These benchmarks can contain different kinds of questions. For example, a dialogue can be about information related to the same entity or entity type.

5.2 Advanced Features

Question understanding: ChatGPT can understand many types of questions better than KGQAN in terms of the linguistic complexity. Benchmarks questions are classified into different types, as proposed in [Dubey *et al.*, 2019]. KGQAN can understand and solve multiple types of questions, such as: (*i*) **single fact:** question about one single piece of information, (*ii*) **single fact with type**: question about a single fact, but the type of the answer is explicitly mentioned in the question, (*iii*) **multi facts:** question asking about the answers using more than one fact, and (*iv*) **boolean:** question giving a fact and asking whether it is true or false.

Other categories that ChatGPT can solve but KGQAN can not include: (*i*) **two intentions:** answers to these questions are more than one fact, e.g. *Who is the wife of Barack Obama and where did he get married?*, (*ii*) **count:** questions asking about the number of answers instead of asking about the answer themselves, e.g.*What is the number of Siblings of Edward III of England ?*. (*iii*) **temporal aspects**: questions asking about facts in a specific time frame, e.g. *With whom did Barack Obama get married in 1992?*

To prove ChatGPT 's ability to understand different questions, we randomly selected a sample of 10 questions per category from LCQuAD-2.0 dataset [Dubey *et al.*, 2019]. For the temporal and two intention questions, ChatGPT managed to understand all of them and answered 90% of the questions correctly. For count questions, ChatGPT did not perform well despite its ability to understand questions. It did not produce any answer for 50% of the questions and managed only to solve correctly 10% of the count questions. KGQAN needs to improve its Seq2Seq model based on the pre-trained language models to support these question types.

Explanability: QASs, such as KGQAN, extract final answers using a structured query fetching information from a KG in a structured format. These answers are not attached with any explanations or references to relevant information. Conversational AI models, such as ChatGPT, provide the user with explanations and additional information about the answer. For QASs, explainability can be in the form of showing the steps followed to produce the answer, i.e. converting the question to the intermediate representation, linking them and getting the answer. This will give the user a better understanding and judgment of the final answer. For example, if the user asks about a specific city, but the systems deal with it as a country, it gives a wrong answer. Without the explanatory steps, the user would not know the reason for the error. But if the system shows the process, the user will clearly understand the reasoning behind the received answer.

User Feedback: There is a need to support different types of user feedback. In the previous example, the user can provide suggestions to improve the quality of the answers, e.g., by clarifying that the question is asking for a city, not a country. In this sense, explainability can lead to having good user feedback and getting the correct answer. Another possible type of user feedback can help maintain or update the information in the KG. In this case, the user may ask a question to extract some information from the KG but notice that this information is outdated. KG chatbot may grant some users privileges to update the KG. Another possibility is that the KG chatbot uses the information given by the user in the context of the current conversation only. After the conversation ends, the system discards the information. Moreover, the KG chatbot could maintain this information as temporary information that needs to be validated before updating the KG.

Identifying Unanswerable Questions: QASs may produce an empty answer for some questions. The user can not know whether the KG does not contain facts related to the question or the system failed to get any answer due to an error in the pipeline. The current benchmarks do not include many unanswerable questions. The KG chatbot should be able to identify unanswerable questions. It may also need to suggest where to find answers as that shows the user the question is understood, but the KG has no related facts.

6 Conclusion

This paper presents a technical survey and comparative framework to formally review the differences between conversational AI models, such as ChatGPT, and traditional question-answering systems (QASs) for knowledge graphs. Our survey conducted a comprehensive evaluation, including four real KGs of different application domains and 450 English questions of various linguistic complexity. Our framework defined seven metrics for quantitative assessment in comparing models, such as ChatGPT and QASs. Our framework identifies the main capabilities required for KG chatbots. Furthermore, our paper highlights open research challenges in advancing and expanding existing QASs with some conversational AI models' abilities, such as dialogue support, robustness, and explainability.

References

[Adiwardana *et al.*, 2020] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, and et al. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977, 2020.

[Auer *et al.*, 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *The International Semantic Web Conference (ISWC)*, volume 4825, pages 722–735, 2007.

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, (EMNLP-IJCNLP)*, pages 3613–3618, 2019.

[Black *et al.*, 2022] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, and et al. Gpt-neox-20b: An open-source autoregressive language model. *CoRR*, abs/2204.06745, 2022.

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, pages 135–146, 2017.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Chowdhery *et al.*, 2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, and et al. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

[CoreNLP, 2022] CoreNLP. https://stanfordnlp.github.io/CoreNLP/, 2022.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pages 4171–4186, 2019.

[Dubey *et al.*, 2019] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *Proceedings of International Semantic Web Conference, (ISWC)*, pages 69–78, 2019.

[Glaese *et al.*, 2022] Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, and et al. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375, 2022.

[Hu *et al.*, 2021] Xixin Hu, Yiheng Shu, Xiang Huang, and Yuzhong Qu. Edg-based question decomposition for complex question answering over knowledge bases. In *Proceedings of the International Semantic Web Conference, (ISWC)*, pages 128–145, 2021.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, and Abdelrahman Mohamed et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 7871–7880, 2020.

[Nakano *et al.*, 2021] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, and et al. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021.

[Omar *et al.*, 2023] Reham Omar, Ishika Dhall, Panos Kalnis, and Essam Mansour. A universal question-answering platform for knowledge graphs. In *Proc. of ACM-SIGMOD*, 2023. Accepted.

[Ouyang *et al.*, 2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, and Carroll L. Wainwright et al. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, and et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, pages 140:1–140:67, 2020.

[Sakor *et al.*, 2019] Ahmad Sakor, Isaiah Mulang, Kuldeep Singh, Saeedeh Shekarpour, and Maria-Esther Vidal et al. Old is gold: Linguistic driven approach for entity and relation linking of short text. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pages 2336–2346, 2019.

[Shin *et al.*, 2020] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, and et al. Biomegatron: Larger biomedical domain language model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 4700–4706, 2020.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 3104–3112, 2014.

[Tanon *et al.*, 2020] Thomas Tanon, Gerhard Weikum, and Fabian Suchanek. YAGO 4: A reason-able knowledge base. In *Proceedings of the European Semantic Web Conference, (ESWC)*, volume 12123, pages 583–596, 2020.

[Taylor *et al.*, 2022] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, and et al. Galactica: A large language model for science. arXiv, 2022.

[Thoppilan *et al.*, 2022] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, and et al. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239, 2022.

[Usbeck *et al.*, 2018] Ricardo Usbeck, Ria Gusmita, Axel-Cyrille Ngomo, and Muhammad Saleem. 9th challenge on question answering over linked data (QALD-9). In *Proceedings of the Workshop on Semantic Deep Learning (SemDeep-4) with the International Semantic Web Conference (ISWC)*, volume 2241, pages 58–64, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, and Llion Jones et al. Attention is all you need. pages 5998–6008, 2017.

[Yan, 2018] Rui Yan. "chitty-chitty-chat bot": Deep learning for conversational AI. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5520–5526, 2018.

[Young *et al.*, 2022] Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of Conference on Artificial Intelligence, (AAAI)*, pages 11622–11629, 2022.

[Zhang *et al.*, 2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, and e al. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022.

[Zhong *et al.*, 2021] Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. Keyword-guided neural conversational model. In *Proceedings of Conference on Artificial Intelligence, (AAAI)*, pages 14568–14576, 2021.