





# Commonsense Reasoning and Explainable Artificial Intelligence Using Large Language Models

Stefanie Krause<sup>(✉)</sup> and Frieder Stolzenburg<sup></sup>

Automation and Computer Sciences Department, Harz University of Applied Sciences, Friedrichstr. 57–59, 38855 Wernigerode, Germany  
{skrause,fstolzenburg}@hs-harz.de  
<http://artint.hs-harz.de/>

**Abstract.** Commonsense reasoning is a difficult task for a computer, but a critical skill for an artificial intelligence (AI). It can enhance the explainability of AI models by enabling them to provide intuitive and human-like explanations for their decisions. This is necessary in many areas but especially in the field of question answering (QA), which is one of the most important tasks of natural language processing (NLP). Over time, a multitude of methods have emerged for solving commonsense reasoning problems such as knowledge-based approaches using formal logic or linguistic analysis.

In this paper, we investigate the effectiveness of large language models (LLMs) on different QA tasks with focus on their abilities on reasoning and producing explanations. For this, we study the recent and very prominent LLM ChatGPT and evaluate the results by means of a questionnaire. We demonstrate ChatGPT’s ability to reason with common sense, and although ChatGPT’s accuracy ranges from 56% to 93% on various QA benchmarks, it outperforms human accuracy. Furthermore we can appraise that, in the sense of explainable artificial intelligence (XAI), ChatGPT gives good explanations for its decisions. In our questionnaire we found that 68% of the participants quantify ChatGPT’s explanations as “good” or “excellent”. Taken together, these findings enrich our understanding of current LLMs and pave the way for future investigations of reasoning and explainability.

**Keywords:** large language models · explainable AI · commonsense reasoning · question answering · ChatGPT

## 1 Introduction

LLMs are an important ingredient in developing adaptable, general language systems [3], and scaling up languages models has recently shown great results for various NLP tasks. Lately, a media hype was triggered by the LLM ChatGPT.<sup>1</sup> This new AI model uses an easy interface and performs very well on

<sup>1</sup> <https://chat.openai.com/>.

different tasks [9]. The current generation of AI systems offers tremendous benefits, but their effectiveness is limited by the inability of the machine to explain its decisions and actions to users. They perceive the models as black boxes although insights about the decision making are mostly opaque [4]. In response to increasing political, ethical, economical, and curiosity-driven theoretical pressure on ML researchers, the field of XAI tries to solve this black-box problem [36]. According to [2], it is important to focus on the audience for which explainability is sought. They define XAI as follows: “Given an audience, an XAI is one that produces details or reasons to make its functioning clear or easy to understand.” They further distinguish between different terms: *explainability* and *interpretability*. Explainability means that a model’s outcome can be explained in human-readable form, e.g., by explanatory text. Interpretability of a model on the other hand refers to the design of the model itself, e.g., so-called heatmaps that visualize neural network activity for image recognition helping to understand the (possibly fallacious) behavior of neural networks [18]. We focus in our work on explainability of AI models in the above sense with the goal of XAI to provide human-readable explanations to make users understand the automated decision-making of large language models a posteriori.

There is a strong connection between XAI and commonsense reasoning, as both concepts are concerned with improving the explainability of AI models. Commonsense reasoning can enhance the explainability of AI models by enabling them to provide intuitive and human-like explanations for their decisions. According to [8], starting with a better understanding of human cognition is a solid foundation. Humans use cognitive reasoning to draw meaningful conclusions despite incomplete and inconsistent knowledge [13]. For us, cognitive reasoning is particularly useful when we encounter new situations that are not covered by formal rules or guidelines. In these situations, we rely on our commonsense to make judgments and decisions that are appropriate and effective. Furthermore, commonsense reasoning is essential in interpersonal interactions and communication. It allows us to understand the perspectives of others and to navigate social situations effectively. Commonsense reasoning can help AI models to be more robust in the context of novel situations. A model that can reason based on commonsense principles is better equipped to handle situations that it has not explicitly encountered before, as it can draw on its general understanding of the world to make informed decisions. So far commonsense reasoning is intuitive for humans but has been a long-term challenge for AI models.

We assume that an LLM can reason similar to humans without the need of logical formulas or explicit ontology knowledge. Recent advances in LLMs (e.g. [22]) have pushed machines closer to human-like understanding capabilities. We believe that language comprehension and commonsense reasoning do not require formal structures, although they eventually may provide a better understanding afterwards for humans. Instead we assume LLMs are the appropriate way towards human-like ability to reason as well as explain decisions. To tackle growing demand of explainability for AI systems we aim to prove that generated explanations by LLMs are helpful for users to understand AI deci-

sions. There is no specific structure of learning necessary: LLMs like ChatGPT can generate human-like explanations a posteriori. For this reason we formulate the following **hypotheses**:

1. LLMs like ChatGPT can handle commonsense reasoning in question answering tasks with near-human-level performance.
2. LLMs like ChatGPT are able to generate good, human-understandable explanations for their decisions.

We start our paper by giving an overview of important research directions in Sect. 2. Then, we evaluate the performance of the recent LLM ChatGPT on commonsense reasoning tasks in Sect. 3. Since measuring explainability is still a problem we address this by first testing ChatGPT on eleven QA datasets where commonsense capabilities are required. With a random sample of each benchmark dataset, subsequently we evaluate the quality of ChatGPT’s responses with a questionnaire (Sect. 4). The main **contributions** of this paper are described in Sect. 5 and can be summarized as following:

- evaluation of ChatGPT’s ability to perform commonsense reasoning
- quality measurement of ChatGPT’s explanations by a questionnaire

## 2 Foundations

### 2.1 Approaches for Commonsense Reasoning

Commonsense reasoning is a difficult task for a computer to handle [32]. To address this problem, various approaches have been followed in the past. McCarthy [23] was the first who outlined the basic approach of representing commonsense knowledge with predicate logic. Symbolic logic approaches were the main representation type, see e.g. [12, 19]. While still in use today [7] for this extremely complex task to work well it requires a large amount of additional logical scaffolding to precisely define the terms used in the statement and their interrelationships [21].

There is a big gap between the logical approach with deductive reasoning and human reasoning, which is largely inductive, associative, and empirical, i.e., based on former experience. Human reasoning, in contrast to formal logical reasoning, does not strictly follow the rules of classical logic. There have been efforts to utilize an approach which uses an automatic theorem prover (that allows to derive new knowledge in an explainable way), large existing ontologies with background knowledge, and recurrent networks with long short-term memory (LSTM) [16] but still did not stand out much from the baseline [32].

Recent efforts to acquire and represent commonsense knowledge resulted in large knowledge graphs, acquired through extractive methods [34] or crowdsourcing [30]. Some approaches use supervised training on a particular knowledge base, e.g., ConceptNet for commonsense knowledge. ConceptNet is a crowd-sourced database that represents commonsense knowledge as a graph of concepts connected by relations [34].

Interestingly, LLMs (cf. Section 2.2) do not contain any explicit semantic knowledge or grammatical let alone logical rules that would allow an explicit reasoning process, not even the large ontologies from the logical knowledge representation like Cyc [19] or Adimen-SUMO [1]. A way out might be to have neural networks learn reasoning explicitly, possibly by focusing on certain sentence forms as in syllogistic reasoning maybe implemented with neural-symbolic cognitive reasoning by specifically structured neural networks [14, 15, 39]. In contrast to simple deep learning, information from different places and/or documents must be merged here in any case. It does not suffice to investigate any local text properties, e.g., determining the text form.

## 2.2 Commonsense Reasoning with LLMs

In the past, most deep learning methods used supervised learning and therefore require substantial amounts of manually labeled data. Recent research has shown that learning good representations in an unsupervised fashion can provide a significant performance boost. An example for a premier LLM that can handle a wide range of natural language processing tasks is OpenAI’s GPT-3 [3]. GPT-3 (Generative Pre-trained Transformer) is a third-generation, autoregressive language model that uses unsupervised learning to produce human-like text. The language model of ChatGPT is trained on an unlabeled dataset of texts, such as Wikipedia to predict the next word for a given text. The capacity of language model is essential to the success of zero-shot task transfer [28]. ChatGPT performs impressive without the need of finetuning on different natural language processing tasks.

The GPT series focuses on pre-training transformer decoders on language modeling. A similar LLM is the Bidirectional Encoder Representations from Transformers (BERT) which uses the transformer encoder as its backbone architecture [10]. BERT obtained new state-of-the-art results on eleven natural language processing tasks already in October 2018 [10]. As well BERT achieved new state-of-the-art performance for example on the SWAG benchmark [38] that exceeded even that of a human expert [5]. However, BERT does not possess human-level commonsense in general [5]. Therefore BERT has been optimized only one year later to RoBERTa to achieve better results [22]. There is also the Bidirectional Auto-Regressive Transformer (BART) [20], a denoising autoencoder for pretraining sequence-to-sequence models, which can be seen as generalizing BERT due to the bidirectional encoder. In our further investigation we will focus solely on ChatGPT. It is a version of GPT with an easy to use interface and at the moment the most prominent LLM.

## 3 Evaluating ChatGPT on QA Tasks

We assess ChatGPT twofold: First, we evaluate the accuracy of ChatGPT on QA benchmarks with multiple-choice questions. In the benchmarks we considered, the correct answer is indicated, although it is not always clear whether this

answer really is the best one. Second, we take part of the questions from the QA benchmarks for a questionnaire to evaluate the quality of the responses and explanations of ChatGPT and compare the performance of humans and ChatGPT on QA examples.

### 3.1 Benchmark Datasets

We use 11 benchmark datasets carefully designed to be difficult to solve without commonsense knowledge (see below). From each dataset, we select 30 random examples, covering different QA tasks like text completion or providing cause or effect. In addition, different fields like medicine, physics, and everyday life situations are covered. We evaluate the performance of ChatGPT with the following QA benchmarks:

- Story Cloze Test [25]: is based on ROC Stories for evaluating story understanding and generation. This test requires choosing the correct ending of a four-sentence story.
- Commonsense Reasoning over Entity Knowledge (CREAK) [26]: contains knowledge about specific entities, e.g., deciding the truthfulness of the claim “Harry Potter can teach classes on how to fly on a broomstick.”, i.e., including fictional worlds. It is bridging fact-checking about entities with commonsense inferences using 13,000 human-authored English claims about entities that are either true or false.
- COMmonsense Dataset Adversarially-authored by Humans (CODAH) [5]: forms a challenging extension to the SWAG dataset [38] which tests commonsense knowledge using sentence-completion questions that describe situations observed in video.
- COM2SENSE [33]: comprises true/false statements, with each sample paired with its complementary counterpart, resulting in 4,000 sentence pairs.
- Cosmos QA [17]: is constructed to test machine reading comprehension with contextual commonsense reasoning. It is a large-scale dataset of 35,600 multiple-choice questions. It focuses on reading between the lines over a diverse collection of people’s everyday narratives.
- Explainable CAusal REasoning dataset (e-CARE) [11]: contains over 21,000 causal reasoning questions, together with natural language formed explanations of the causal questions.
- AI2 Reasoning Challenge (ARC) [6]: covers natural, grade-school science questions that are authored for human tests, and is the largest public-domain set of this kind with 7,787 questions.
- Social IQa [31]: contains 38,000 multiple choice questions for probing emotional and social intelligence in a variety of everyday situations.
- Choice Of Plausible Alternatives (COPA) [29]: is an open-domain commonsense reasoning QA task where each question gives a premise and two plausible causes or effects, where the correct choice is the alternative that is more plausible than the other. The cause category requires backward causal reasoning, while the result category requires forward causal reasoning.

- Multiple-Choice Question Answering (MedMCQA) [27]: is designed to address real-world medical entrance exam questions. It is covering more than 2,400 healthcare topics and 21 medical subjects.
- CommonsenseQA [35]: for this dataset crowd-workers authored multiple-choice questions that mention the source concept extract from ConceptNet [34] multiple target concepts that have the same semantic relation to a single source concept.

### 3.2 Method


Using all these eleven datasets we randomly selected 30 examples from each dataset and tested the respective QA tasks with ChatGPT. Over all datasets ChatGPT answered 242 out of 330 tasks correctly which equals an accuracy of 73.33%, 77 tasks were answered incorrectly (23.33%), and we did not get a valid response for 11 QA tasks (3.33%). Not valid means that ChatGPT does not respond which answer option is correct and instead asks for further context information, see Fig. 1 for an example. Note that there are tasks with two answer options and tasks with five answer options and everything in between.

SK

The family searched the neighborhood. What is the cause?

1. Their dog ran away from home.
2. Expensive jewelry was missing from their home.

Which alternative is the more likely cause?



Based on the information provided, we cannot determine the more likely cause between the two alternatives. The sentence "The family searched the neighborhood" doesn't give enough context to make a definitive judgment.

Both alternatives are possible causes for the family searching the neighborhood. It's equally plausible that the family searched for their dog, which might have run away from home, or they searched for expensive jewelry that was missing from their home.

Without additional information or context, we cannot determine which alternative is the more likely cause.

**Fig. 1.** Example for a not valid response from ChatGPT due to insufficient context information (COPA example 612). While option 1 is correct, the authors agree that option 2 could also be possible as well (but less likely).

A more detailed representation of the performance on each of the eleven datasets is shown in Table 1. We found that ChatGPT has the lowest performance on CommonsenseQA dataset with 56.67% accuracy and the highest accuracy on Story Cloze Test with 93.33%.

**Table 1.** Overview of eleven publicly available datasets for commonsense reasoning. For each dataset we report the year the dataset was published and the percentage of correct, incorrect and invalid answers of ChatGPT on 30 randomly selected examples per dataset.

dataset	year	correct	incorrect	invalid
Story Cloze Test [25]	2017	93.33%	6.67%	0.00%
CREAK [26]	2021	86.67%	13.33%	0.00%
CODAH [5]	2019	80.00%	20.00%	0.00%
COM2SENSE [33]	2021	76.67%	23.33%	0.00%
CosmosQA [17]	2019	76.67%	23.33%	0.00%
e-CARE [11]	2022	76.67%	23.33%	0.00%
ARC [6]	2018	70.00%	30.00%	0.00%
Social IQa [31]	2019	66.67%	33.33%	0.00%
COPA [29]	2011	63.33%	3.33%	33.33%
MedMCQA [27]	2022	60.00%	40.00%	0.00%
CommonsenseQA [35]	2018	56.67%	43.33%	0.00%

### 3.3 Analysis

In our error analysis we found that there are six kinds of problems where ChatGPT still struggles:

1. **missing context:** In cases where ChatGPT has little knowledge of the context provided, it sometimes does not give an answer to the QA task. This has happened 10 times in total and solely with examples of the COPA dataset. This could be due to the very short premise texts in the COPA dataset, see Fig. 1. In this dataset the premise texts consist of only five to nine words (on average six words) in the cases where ChatGPT complained about not having enough information to answer the question. In some cases, ChatGPT explains which context information is missing: “The actual outcome would depend on a variety of factors, such as the political climate, the credibility of the politician, and the specific details of the argument in question. Without this information it is impossible to determine which alternative is more likely.” (COPA example 619).
2. **comparative reasoning:** ChatGPT has problems when more than one option is plausible. This is the case in comparative scenarios in the COM2SENSE and Social IQa dataset. In such cases, the commonsense reasoner must explicitly investigate the likelihood of different answer candidates. For the Social IQa example 26823 “Sasha was throwing a party in her new condo which they bought a month ago. What does Sasha need to do before this?” ChatGPT answers “Turn music on” which is likely but the correct and even more likely answer is “needed to buy food for the party”.
3. **subjective reasoning:** Some answers depend on the personality of the reasoner, e.g. Social IQa example 18571: “Alex’s powers were not as strong since

he was just starting out. Alex used Bailey’s powers since hers were stronger. How would Bailey feel as a result?” the correct answer according to the benchmark is “good” but instead ChatGPT answers “upset” and explains “Bailey may feel that her powers are being taken advantage of . . .” which we think is more a personalized subjective inference instead of a commonsense answer.

4. **slang, unofficial abbreviations, and youth language:** ChatGPT has its difficulties to understand slang, unofficial abbreviations and youth language like “subs” for “subscribers” or “yrs” for “years”. This could be observed in Cosmos QA examples 6599 and 5748.
5. **social situations:** We identified a lack of understanding social situations correctly especially in the Social IQa dataset. For example, for the question “Kai was visiting from out of state and brought gifts for Quinn’s family. What will Kai want to do next?” ChatGPT picked the answer “needed to leave his hometown” instead of the correct answer option “watch the opening of gifts” (Social IQa example 6863).
6. **medical domain:** The analysis of MedMCQA showed that ChatGPT is lacking a deep domain knowledge in the medical field. The answers of ChatGPT were always plausible and explained with a lot of details (on average 43 words per explanation) but 40% were incorrect. This was because of many medical technical terms that are not common knowledge, e.g., “Styloglossus muscle” or “Genioglossus muscle” that are different muscles in the tongue (MedMCQA example 23b363d6-8210-4657-b293-54c9e28bdf31). For a non-medical professional or student, these questions are difficult to answer, too (including the authors of this paper).

Please be aware that for certain questions to be answered correctly, one must possess in-depth knowledge rather than commonsense reasoning ability, e.g., you have to know that “Prison Break” is a television show, not a movie in a theater to tell that “The couple went to the movie theater to watch Prison Break” is a correct or wrong statement (CREAK example 98). Additionally, the authors hold the viewpoint that out of the 78 incorrect answers, 12 of them were very likely to be correct as well and therefore quite hard for an AI to answer correctly.

### 3.4 Design of the Questionnaire

To evaluate the quality of ChatGPT’s responses on different benchmark datasets and to make a comparison to human performance, we created a questionnaire. We used two randomly selected examples for each of the above mentioned datasets – except for MedMCQA because we feel these questions are too difficult for non-medical people to answer.

We created an online survey questionnaire using SoSciSurvey<sup>2</sup> that was open to the public on social media, e.g., LinkedIn, Xing, and platforms like Survey-Circle and we send the questionnaire via e-mail directly to students at the Harz University of Applied Sciences in Germany. Participation was voluntary; participants could not be identified from the material presented and no plausible harm

<sup>2</sup> <https://www.sosicisurvey.de/>.



to participating individuals could arise from the study. Survey content validity was reviewed in a pretest by one professor, one academic staff and one non-academic volunteer (business consultant) who did not participate in developing the survey. The questionnaire was structured in three parts, first containing demographic and personal information (gender, age, nationality, English level). The main part then consists of the QA tasks of the different datasets as well as an evaluation of ChatGPT’s explanations. For each QA task we have the same structure, as follows:

1. The question and answer options for each QA task were given for the survey participants.
2. We ask how comprehensible the question above is using a five-level Likert scale.
3. The question and answer options were repeated and ChatGPT’s explanation for one possible answer option is presented (this answer may be incorrect). Then using a five-level Likert scale we ask how good the explanation is.
4. An optional free text field to tell what could be improved in the given explanation.

To see an example of this main questionnaire section, refer to Fig. 2. We used this structure for 10 datasets and randomly selected two examples from every dataset. Therefore we considered  $2 \cdot 10 = 20$  QA tasks. In the third part of the questionnaire, the participants should guess how many explanations have been generated by an AI among others. Note that the survey participants did not know that *all* responses have been generated by ChatGPT.

## 4 Results

### 4.1 Questionnaire Participants

In total, 103 people participated in the questionnaire, but because of missing data we only used the responses of 49 participants. The time to fully answer the whole questionnaire was about 25 min, that is probably why many participants did not complete the questionnaire until the last question. The participants English level was mainly advanced or excellent so that there is no language barrier in understanding the QA tasks. Among the completed questionnaires, 71% of the participants were male and the average age was 26 years, with a minimum of 19 years and maximum of 49 years. Most of the participants were German with 45% and Indian with 40% and only 5% of Bangladeshi, Pakistan, Finland, Russian Federation and Switzerland.

### 4.2 Questionnaire Responses

We found that the participants answered 73.72% of the 20 QA tasks correctly compared to ChatGPT’s 90.00% on the same questions. Note that these 20 in detail analyzed QA tasks are not as representative as the 330 QA tasks

**1. Is this correct?**  
 William Howard Taft served in the United States government.

yes, correct  
 no, incorrect

**2. How comprehensible do you find the question?**

very poor  
 poor  
 fair  
 good  
 excellent

**Is this correct?**  
 William Howard Taft served in the United States government.

yes, correct  
 no, incorrect

**Possible response: yes, correct**

**Explanation: Howard Taft served as the 27th President of the United States from 1909 to 1913. Prior to his Presidency, Taft served as the Secretary of War from 1904 to 1908.**

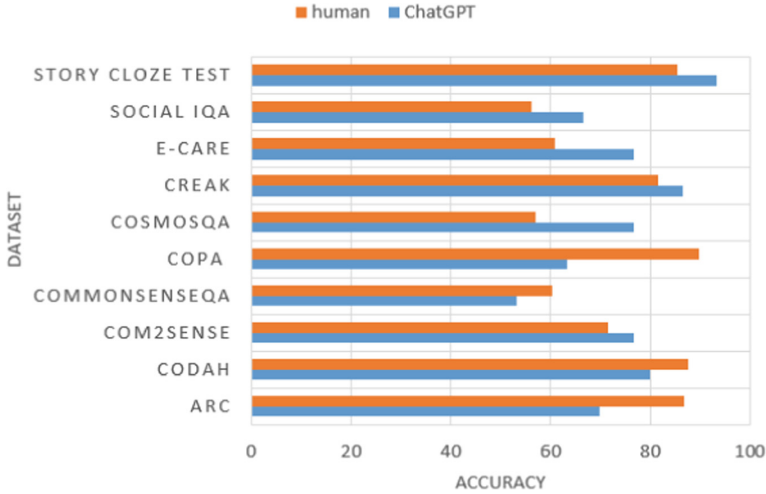
**3. How good is the explanation?**

very poor  
 poor  
 fair  
 good  
 excellent

**(optional) What could be improved in the given explanation?**

**Fig. 2.** Example of one question answering task with the common structure of four questions per task (CREAK example 1344). After participants answer the first two questions the next two questions with the possible response and explanation are shown.

from Sect. 3.2. Even though we selected the 20 QA tasks randomly, ChatGPT performed much better on these subset than on the overall set of QA tasks. Over all datasets, except MedMCQA, ChatGPT answered 74.67% correct of the  $30 \cdot 10 = 300$  tasks. Figure 3 shows a comparison of the performance of ChatGPT and the survey participants on the different datasets. The performance of Chat-

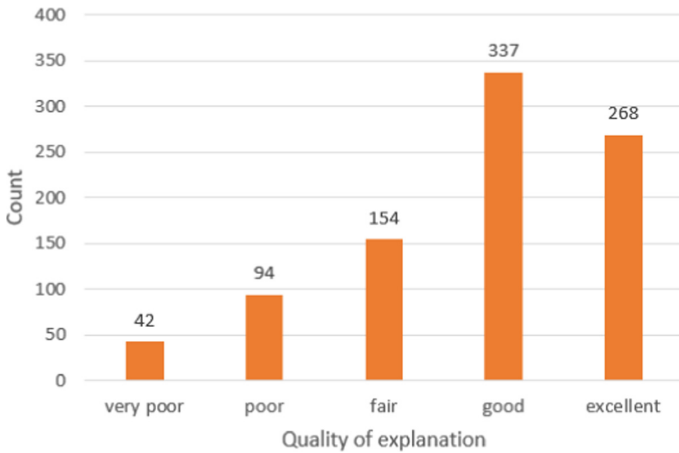


**Fig. 3.** Comparison of accuracy of ChatGPT (blue) and our survey participants (orange) on ten different QA datasets. (Color figure online)

GPT is better than humans on six datasets and in four datasets humans are superior. The dataset ChatGPT performs worst is also the dataset humans had the most problems (CommonsenseQA). The greatest difference between human and ChatGPT performance is on the COPA and Cosmos QA datasets. In this study, for COPA examples, humans were 26.47% better than ChatGPT and, for the Cosmos QA examples, ChatGPT outperforms humans with 19.53% difference in accuracy. It is quite interesting that ChatGPT performs better on Cosmos QA than the survey participants as contextual commonsense reasoning is needed for this dataset. It focuses on reading between the lines over a diverse collection of people’s everyday narratives. In contrast, humans perform a lot better than ChatGPT on COPA where understanding of causes and effects is necessary as well as choosing the most likely alternative. Our study showed that ChatGPT has problems with comparative reasoning in case of more than one likely option. Maybe here explicit traditional reasoning approaches from AI maybe would perform better (cf. Section 1).

We were interested in investigating the relationships between tasks comprehensibility and ChatGPT’s explanations. It is worth noting that most questions of the different QA tasks are comprehensible according to the participants. We observed that there is a mean linear positive correlation of 0.58 between the comprehensibility of the QA tasks and that of ChatGPT’s explanations. This means that the way the users understand the QA tasks has an impact on the estimated quality of the explanation from ChatGPT. The Social IQa examples 23772 and 11339 were rated 22 times out of 56 total times as very poorly comprehensible. Nevertheless, ChatGPT answered these tasks correctly but only 56.13% survey participants answered these questions correctly.

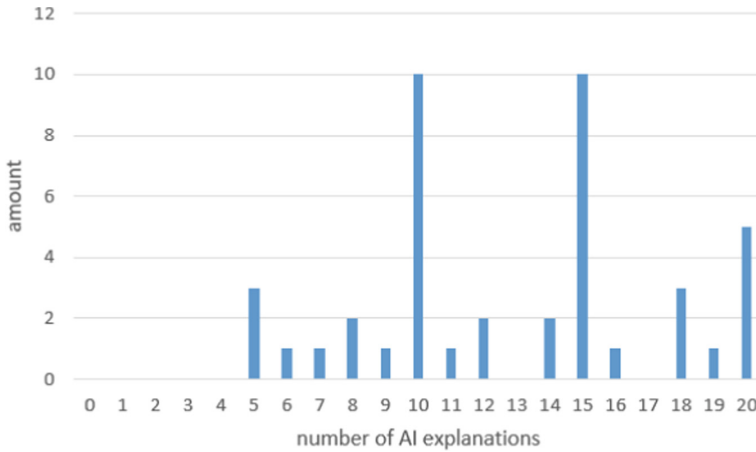
Furthermore, we found that the explanations for COPA example 610 was often rated “poor” or “very poor”, and for this example ChatGPT’s answer was invalid as it could not decide for one option saying: “It is not specified in the given information which alternative is the more likely cause.” In general, explanations were mostly rated “good” or “excellent” with 67.60% and only 42 times very poor (see Fig. 4). Explanations were rated “fair” or better with 84.80%. In this study, 12 out of 42 times the explanations were rated “very poor” for the undefined responses, where ChatGPT was unable to answer the question due to missing context information (see example above in Fig. 1). The average length of ChatGPT’s explanations is 38 words for both correct and incorrect responses. From the optional free text field we received mostly the same possible improvement for ChatGPT’s explanation: ChatGPT should explain why the other answering options are false or less likely and not only focus on explaining why one option is correct. This is in particular important in comparative reasoning tasks.



**Fig. 4.** Participants’ rating of all explanations from “very poor” to “excellent”.

Figure 5 shows the participants’ guess how many responses are created by an AI according to the survey participants. In the chart one can see that the mode is 10 and 15 explanations. While all respondents think that at least five explanations are generated by an AI, the mean amount of AI answers is 13. Thus all participants believe that at least 25% of the explanations were AI generated.

To further determine how helpful explanations are, we ask our study participants if they agree that AI tools that give not only a decision but also an explanation should be preferred. The majority (52.08%) of the participants agreed to this statement, 31.25% agree strongly and 10.42% are neutral while less than 7% disagree or strongly disagree.



**Fig. 5.** Bar chart of participants option how many explanations are generated by an AI. The average number of assumed AI answers is 13 while actually all 20 explanations were generated by ChatGPT.

## 5 Discussion and Future Directions

Over the past, research often focused on logical approaches and large knowledge graphs to deal with commonsense reasoning. Given that we are currently in the era of LLMs which have shown substantial performance improvements across various tasks, we hypothesized that LLMs are capable of handling commonsense reasoning in QA tasks with almost human-level performance (Hypothesis 1). As ChatGPT is trained on a large number of data and produces human-like text, we assume that it can perform commonsense reasoning without explicit semantic knowledge or logical rules. To proof that we evaluated ChatGPT on eleven different QA benchmark datasets which are difficult to solve without commonsense reasoning.

Moreover we evaluated explanations generated with ChatGPT by means on an online questionnaire to investigate how sufficient explanations are to users. Our Hypothesis 2 is that an LLM like ChatGPT is able to provide good explanations to users without the need of explicit formalized knowledge representation. Most participants are content with ChatGPT’s explanations. Thereby apparently the problem of explainability of AI decisions can be overcome easily.

### 5.1 Main Findings

This study shows that ChatGPT reached an overall accuracy of 73.33% on eleven QA datasets that are difficult to handle without commonsense reasoning. While there are still problems (cf. Section 3.3), ChatGPT still outperforms our survey participants in six out of ten datasets (not considering the medical dataset MedMCQA). The results of our questionnaire show that participants answered

73.72% of the 20 QA tasks correctly compared to 90.00% of ChatGPT on the same questions. Although we only compared performance of humans vs. ChatGPT on a small amount of examples, we beforehand evaluated ChatGPT on eleven different benchmarks on a larger set of examples. Consequently, we believe that the outcome indicates that our Hypothesis 1 is true and LLMs like ChatGPT can handle commonsense reasoning in QA tasks with near-human-level performance.

This research focused as well on assessing explainability of LLMs, recognizing the significant importance of addressing the black-box problem. This is particularly relevant as users need to understand AI decisions. By means of a web-based questionnaire we evaluated ChatGPT’s explanations for 20 QA tasks. We found a mean linear positive correlation of 0.58 between the comprehensibility of the QA tasks and that of ChatGPT’s explanations. This observation is relevant for the way ChatGPT’s users describe their tasks as it has an impact on the quality of the explanation they receive. In our questionnaire, ChatGPT’s explanations were mostly rated “good” or “excellent” with 67.60%. Our Hypothesis 2 that LLMs can generate good explanations could be confirmed. However, to improve explanations, it is recommended to not only focus on explaining why one option is correct but also why the other answering options are false or less likely.

## 5.2 Impact on the Field

The development of XAI is facing both scientific and social demands [37], and scientists aim to achieve this without sacrificing performance. So far, this grand challenge is mainly dealt by explicit knowledge, such as knowledge graphs. However, we found that implicit knowledge in the form of probabilistic models can generate good explanations. LLMs, such as GPT, made significant advancements in NLP tasks in recent years. Due to the chat function of ChatGPT, users can easily ask for explanations to understand the response of the AI system. This can tackle the lack of explainability and is a quite simple and yet effective way. Using a questionnaire, we could measure and quantify explanations of ChatGPT and investigate the effectiveness of AI explanations.

Moreover, commonsense reasoning is very important for various NLP tasks. It assesses the relative plausibility of different scenarios and recognizes causality. Until now, research focuses on mathematical logic and the formalization of commonsense reasoning knowledge. However, some philosophers, e.g., Wittgenstein, already claimed that commonsense reasoning knowledge is unformalizable or mathematical logic is inappropriate [24]. As seen in our evaluation, the LLM ChatGPT can handle different QA tasks that require commonsense reasoning. Nevertheless, we detected six problems (cf. Section 3.3) where ChatGPT has still problems and further research is necessary. These difficulties are little context information, comparative reasoning, subjective reasoning, slang, unofficial abbreviations and youth language, social situations and knowledge in the medical domain.

Evaluation of the LLM ChatGPT brings AI closer to making a practical impact in the area of XAI and commonsense reasoning. There are still rich

opportunities for novel AI research to further measure the quality of explanations as well as opportunities in tackling difficult commonsense reasoning tasks like CommonsenseQA. In future research, one can also investigate other LLMs than ChatGPT, e.g., BERT, BART, RoBERTa, etc.

### 5.3 Limitations

Our study has limitations that need to be acknowledged. The number of survey participants we included was rather small, which limits generalization of our results. The average age was 26 years with 49 years as maximum, and primarily the participants were university students. In general more participants with diverse gender, age and nationality would help to strengthen the results. Furthermore the key challenge for explainability is to determine what constitutes a “good” explanation, since this is subjective and depends on context. We evaluated explanations using a five-level Likert scale from “very poor” to “excellent”. However, we only analyzed 20 explanations of ChatGPT and argue that our Hypothesis 2 (that LLMs can generate good explanations) can be confirmed. Nevertheless, explainability is very important in the medical field, but we did not consider the MedMCQA dataset in our questionnaire due to a supposed lack of participants knowledge in medicine.

## 6 Conclusion

The field of AI has made considerable progress towards large-scale models, especially for NLP tasks. Although the field requires more testing, we argue that LLMs can be used for commonsense reasoning tasks and as well generate helpful explanations for users to understand AI decisions. The use of LLMs is a promising area of research that offers many opportunities to enhance explainability. However, to unleash their full potential for XAI, it is crucial to approach the use of these models with caution and to critically evaluate their limitations. We have shown important future directions and rich opportunities for novel AI research involving XAI and commonsense reasoning. LLMs have proven capable of human-like performance on a variety of different QA tasks which require commonsense reasoning.

Despite the potential of the field of LLMs, important questions remain for a comprehensive evaluation of ChatGPT’s explanations. As these key issues are systematically addressed, the potential of AI to significantly improve the future of XAI may be realized. In particular, the stochastic aspects of LLMs, where repeated queries may lead to different answers, should be considered in future work. This would also allow for a better assessment of the error in the ChatGPT performance estimates.

**Acknowledgments.** We would like to thank Oliver Obst for sharing our questionnaire on LinkedIn and Osama Siddiqui for his help testing ChatGPT on different datasets. We would also like to thank the anonymous reviewers for their thoughtful reading and comments.

## References

1. Álvarez, J., Lucio, P., Rigau, G.: Adimen-SUMO: Reengineering an ontology for first-order reasoning. *Int. J. Semant. Web Int. Syst. (IJSWIS)* 8(4), 80–116 (2012). <https://ideas.repec.org/a/igg/jswis0/v8y2012i4p80-116.html>
2. Barredo Arrieta, A., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020). <https://doi.org/10.1016/j.inffus.2019.12.012>
3. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 1877–1901 (2020). <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
4. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.* **70**, 245–317 (2021). <https://doi.org/10.1613/jair.1.12228>
5. Chen, M., D’arcy, M., Liu, A., Fernandez, J., Downey, D.: CODAH: An adversarially-authored question answering dataset for common sense. In: *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pp. 63–69 (2019). <https://www.jaredfern.com/publication/codah/>
6. Clark, P., et al.: Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR - Computing Research Repository* abs/1803.05457, Cornell University Library (2018), <https://arxiv.org/abs/1803.05457>
7. Davis, E.: Logical formalizations of commonsense reasoning: a survey. *J. Artif. Intell. Res.* **59**, 651–723 (2017). <https://doi.org/10.1613/jair.5339>
8. Davis, E., Marcus, G.: Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* **58**(9), 92–103 (2015). <https://doi.org/10.1145/2701413>
9. Deng, J., Lin, Y.: The benefits and challenges of ChatGPT: An overview. *Front. Comput. Intell. Syst.* 2(2), 81–83 (2023). <https://doi.org/10.54097/fcis.v2i2.4465>
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. <https://arxiv.org/pdf/1810.04805>
11. Du, L., Ding, X., Xiong, K., Liu, T., Qin, B.: e-CARE: a new dataset for exploring explainable causal reasoning. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 432–446. Association for Computational Linguistics (2022). <https://aclanthology.org/2022.acl-long.33/>
12. Forbus, K.D.: Qualitative process theory. *Artif. Intell.* **24**(1–3), 85–168 (1984). [https://doi.org/10.1016/0004-3702\(84\)90038-9](https://doi.org/10.1016/0004-3702(84)90038-9)
13. Furbach, U., Hölldobler, S., Ragni, M., Schon, C., Stolzenburg, F.: Cognitive reasoning: A personal view. *KI* 33(3), 209–217 (2019). <https://link.springer.com/article/10.1007/s13218-019-00603-3>
14. d’Avila Garcez, A.S., Broda, K., Gabbay, D.M.: Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence* 125(1–2), 155–207 (2001). [https://doi.org/10.1016/S0004-3702\(00\)00077-1](https://doi.org/10.1016/S0004-3702(00)00077-1)
15. d’Avila Garcez, A., Lamb, L., Gabbay, D.: *Neural-Symbolic Cognitive Reasoning*. Springer, Berlin, Heidelberg (2009). <https://doi.org/10.1007/978-3-540-73246-4>
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>



17. Huang, L., Le Bras, R., Bhagavatula, C., Choi, Y.: Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2391–2401. Association for Computational Linguistics (2019). <https://aclanthology.org/D19-1243/>
18. Lapuschkin, S., Waldchen, S., Binder, A., Montavon, G., Samek, W., Muller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications* 10(1), 1096 (2019). <https://www.nature.com/articles/s41467-019-08987-4>
19. Lenat, D.B.: CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM* 38(11), 33–38 (1995). <https://doi.org/10.1145/219717.219745>
20. Lewis, M., et al.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019). [arXiv:1910.13461](https://arxiv.org/abs/1910.13461)
21. Liu, Hugo, Singh, Push: Commonsense Reasoning in and Over Natural Language. In: Negroita, Mircea Gh., Howlett, Robert J., Jain, Lakhmi C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 293–306. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30134-9\\_40](https://doi.org/10.1007/978-3-540-30134-9_40)
22. Liu, Y., et al: RoBERTa: A robustly optimized BERT pretraining approach. <https://arxiv.org/pdf/1907.11692>
23. McCarthy, J.: Programs with common sense (1959). <https://www.cs.rit.edu/~rlaz/is2014/files/McCarthyProgramsWithCommonSense.pdf>
24. McCarthy, J.: Artificial intelligence, logic and formalizing common sense. In: Thomason, R.H. (ed.) *Philosophical Logic and Artificial Intelligence*, pp. 161–190. Springer, Netherlands, Dordrecht (1989). [https://doi.org/10.1007/978-94-009-2448-2\\_6](https://doi.org/10.1007/978-94-009-2448-2_6)
25. Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., Allen, J.: LSDSem 2017 shared task: The Story Cloze Test. In: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pp. 46–51 (2017). <https://aclanthology.org/W17-0906.pdf>
26. Onoe, Y., Zhang, M.J.Q., Choi, E., Durrett, G.: CREAK: A dataset for commonsense reasoning over entity knowledge. <https://arxiv.org/pdf/2109.01653>
27. Pal, A., Umaphathi, L.K., Sankarasubbu, M.: MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *ACM Conference on Health* (2022). <https://arxiv.org/pdf/2203.14371>
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. Tech. rep., OpenAI (2019). <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
29. Roemmele, M., Bejan, C.A., Gordon, A.S.: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In: *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pp. 90–95 (2011), <https://aaai.org/papers/02418-choice-of-plausible-alternatives-an-evaluation-of-commonsense-causal-reasoning/>
30. Sap, M., et al.: ATOMIC: an atlas of machine commonsense for if-then reasoning. *Proc. AAAI Conf. Artif. Intell.* 33(01), 3027–3035 (2019). <https://doi.org/10.1609/aaai.v33i01.33013027>
31. Sap, M., Rashkin, H., Chen, D., LeBras, R., Choi, Y.: Social IQa: Commonsense reasoning about social interactions. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4463–4473.

- Association for Computational Linguistics (2019). <https://aclanthology.org/D19-1454/>
32. Siebert, S., Schon, C., Stolzenburg, F.: Commonsense reasoning using theorem proving and machine learning. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) *Machine Learning and Knowledge Extraction - 3rd IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019*, pp. 395–413. LNCS 11713, Springer Nature Switzerland, Canterbury, UK (2019). [https://doi.org/10.1007/978-3-030-29726-8\\_25](https://doi.org/10.1007/978-3-030-29726-8_25)
  33. Singh, S., et al.: COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 883–898. Association for Computational Linguistics (2021). <https://aclanthology.org/2021.findings-acl.78>
  34. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1) (2017). <https://doi.org/10.1609/aaai.v31i1.11164>
  35. Talmor, A., Herzig, J., Lourie, N., Berant, J.: CommonsenseQA: A question answering challenge targeting commonsense knowledge. <https://arxiv.org/pdf/1811.00937>
  36. Taylor, J.E.T., Taylor, G.W.: Artificial cognition: how experimental psychology can help generate explainable artificial intelligence. *Psychon. Bull. Rev.* 28(2), 454–475 (2021). <https://doi.org/10.3758/s13423-020-01825-5>
  37. Xu, Feiyu, Uszkoreit, Hans, Du, Yangzhou, Fan, Wei, Zhao, Dongyan, Zhu, Jun: Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In: Tang, Jie, Kan, Min-Yen., Zhao, Dongyan, Li, Sujian, Zan, Hongying (eds.) *NLPCC 2019. LNCS (LNAI)*, vol. 11839, pp. 563–574. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32236-6\\_51](https://doi.org/10.1007/978-3-030-32236-6_51)
  38. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y.: SWAG: A large-scale adversarial dataset for grounded commonsense inference. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93–104. Association for Computational Linguistics (2018). <https://aclanthology.org/D18-1009/>
  39. Zimmer, M., et al.: Differentiable logic machines. *CoRR - Computing Research Repository* abs/2102.11529, Cornell University Library (2021). <https://arxiv.org/abs/2102.11529>, latest revision 2023