Desiderata for the Context Use of Question Answering Systems

Sagi Shaier,[◇] Lawrence E Hunter,[†] Katharina von der Wense[▽◇]
[▽]University of Colorado Boulder
[†]Independent Scholar
[◇]Johannes Gutenberg University Mainz
[▽]E-mail: {sagi.shaier, katharina.kann}@colorado.edu
[†]E-mail: Prof.Larry.Hunter@gmail.com

Abstract

Prior work has uncovered a set of common problems in state-of-the-art context-based question answering (QA) systems: a lack of attention to the context when the latter conflicts with a model's parametric knowledge, little robustness to noise, and a lack of consistency with their answers. However, most prior work focus on one or two of those problems in isolation, which makes it difficult to see trends across them. We aim to close this gap, by first outlining a set of – previously discussed as well as novel – desiderata for QA models. We then survey relevant analysis and methods papers to provide an overview of the state of the field. The second part of our work presents experiments where we evaluate 15 QA systems on 5 datasets according to all desiderata *at once*. We find many novel trends, including (1) systems that are less susceptible to noise are not necessarily more consistent with their answers when given irrelevant context; (2) most systems that are more susceptible to noise are more likely to correctly answer according to a context that conflicts with their parametric knowledge; and (3) the combination of conflicting knowledge and noise can reduce system performance by up to 96%. As such, our desiderata help increase our understanding of how these models work and reveal potential avenues for improvements. Code and data can be found here: https://github.com/Shaier/context_usage_desiderata.git.

1 Introduction

Question answering (QA) systems which are based on large language models (LLMs) play a larger role than ever before in our society, due to their ability to offer quick access to information (Petroni et al., 2019; Roberts et al., 2020; Shin et al., 2020; Sung et al., 2021; Jiang et al., 2020a). Many QA systems can make use of context information when available, which often contains relevant information to help systems answer questions, cf. Figure 1. We refer to all systems that are able to leverage such context information as *context-based QA systems*.

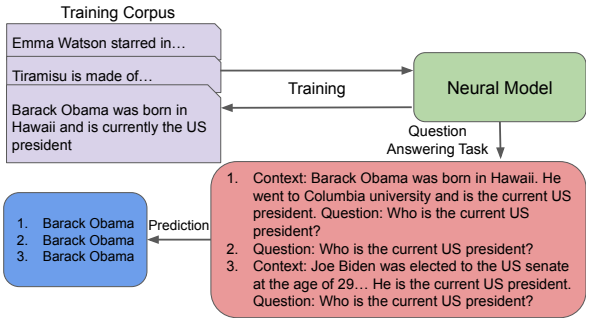

Figure 1: An example where the model was trained to learn the knowledge "Barack Obama is the current US president". In the first and second tasks the model answered the questions correctly. However, in the third task where the model is given context with conflicting information it fails to answer the question correctly.

Many aspects of such systems have been evaluated by previous work, such as the amount of their parametric knowledge (Petroni et al., 2019) and their robustness to noise (Jia and Liang, 2017), conflicting knowledge (Pan et al., 2021; Longpre et al., 2021), or irrelevant contexts (Li et al., 2022; Neeman et al., 2022) . However, looking at such aspects in isolation makes it difficult to see trends across problems, e.g., to explore whether there is a connection between a model's attention to context and its ability to handle noise.

Here, we 1) outline a set of – previously discussed as well as novel – desiderata for context-based QA models and 2) provide an extensive survey of related works, which we group and discuss according to our desiderata. Such desiderata unify some of the existing aspects from the literature, e.g., robustness to conflicting knowledge, and outline how a QA system should behave from the perspective of the context. We will publicly release a toolkit to prepare datasets — both free-form and multiple choice (MC) type – to evaluate models

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 777–792

	Context	Distractor	Known Knowledge	Unknown Knowledge
1	Original		T	T
2	Original	✓	T	T
3	Alternative		A	A
4	Alternative	✓	A	A
5	None		T	B
6	Irrelevant		T	B

Table 1: Desiderata table: what should an optimal model do for different types of contexts? T = *true answer*; A = *conflicting answer*; B = *wrong answer/unanswerable*; The distractor (cf. Sec. 3.3) is a string of words that is concatenated to the context (cf. Sec. 4.1.3); Alternative context (cf. Sec. 3.4) is a slight modification of the original context, where we replace the answer string with an alternative one (cf. Sec. 4.1.1); Irrelevant context (cf. Sec. 3.6) is a random context (cf. Sec. 4.1.2).

according to all desiderata *at once*.

Using our toolkit, we 3) evaluate 15 LLM-based QA systems and first confirm prior works' results: while some systems appear nearly perfect, scoring 99% accuracy on standard datasets, their performance is significantly worse according to many of our desiderata. For instance, their accuracy drops by up to 93% with noise, such as random strings as distractors. Second, considering all desiderata *at once*, we find that (1) systems that are less susceptible to noise are more consistent with their answers when provided with irrelevant context; (2) most systems that are more susceptible to noise are more likely to correctly answer according to a context that conflicts with their parametric knowledge; and (3) the combination of conflicting knowledge and noise can reduce system performance by up to 96%. Finding these novel trends using our desiderata opens new avenues to improve QA models.

2 Desiderata

We now develop a set of desiderata regarding the context use of a model, before presenting our survey on what prior work has found with regards to our desiderata and performing our own experiments in the next two sections. To come up with our desiderata, which are presented in Table 1, we consider the question: *How would an ideal QA model behave for different types of context?*

The ideal behavior depends on whether the knowledge in the context is known or unknown to the model. For example, looking at Table 1, Row 6, while systems are expected to predict the true answer for known knowledge, as they contain the relevant context within their parameters, the ideal system would answer incorrectly/"unanswerable" for unknown knowledge given irrelevant context.

We follow the work by Li et al. (2022); Xie et al. (2023); Roberts et al. (2020) and define known knowledge as questions that a model can answer correctly without context, and unknown knowledge as those it cannot.

Proposed Desiderata An ideal model should:

■ For both known and unknown knowledge:

a. *Answer correctly with the original context*: this is the standard QA systems evaluation approach.

b. *Answer correctly with a noisy irrelevant variation of the original context*: QA systems should be robust to distractors, as different users and information retrieval (IR) systems introduce varying amounts of irrelevant information.

c. *Change its answer with conflicting context to the conflicting knowledge*: As our world is constantly changing, QA systems should be dynamic in their knowledge. That is, similarly to Zhou et al. (2023); Li et al. (2022), we believe that the context should *always* take priority over a model's parametric knowledge, when relevant.

■ For known knowledge:

d. *Answer correctly with no context*: In our setting this happens by default for known knowledge, as by definition known knowledge is questions that can be answered without context. However, we expect the ideal system to have the largest possible amount of knowledge, i.e., to be able to answer most questions without context.

e. *Answer correctly with an irrelevant context*: Since the model answers questions correctly without context for known knowledge, it should also answer correctly with irrelevant context.

■ For unknown knowledge:

f. *Answer incorrectly/"unanswerable" with no context*: In our setting this happens by default for unknown knowledge. *While the ideal model should predict "unanswerable" for questions it cannot answer*, most existing datasets do not include questions that, according to our definition, are truly "unanswerable," as they *can* be answered with parametric knowledge (cf. Sec. 4.1.2). Hence, we add here that models may also predict an incorrect answer, as expected from models that are forced to predict any answer other than "unanswerable" for unknown knowledge.

g. *Answer the same with irrelevant context as with*

no context: The ideal model should be consistent in its answer, even when wrong. Hence, the model's answer with irrelevant context and no context – (f) above – should be the same.

3 Survey of Prior Work

3.1 Known vs. Unknown Knowledge

As mentioned in Sec. 2, the ideal behavior depends on if the knowledge contained in the context is known or unknown to the model. While most work evaluate on the entire data without such distinction, some analyze the known knowledge split: Xie et al. (2023); Li et al. (2022) analyze models using a closed-book setting, Neeman et al. (2022) assume the original contexts are known knowledge, and Chen et al. (2022) evaluate correctly answered questions.

3.2 The Standard Approach

Row 1 in Table 1 shows the standard approach for evaluating QA systems, where systems are tasked with answering questions using a fixed context. For lack of space and since the focus of our survey is not the standard approach, we refer interesting readers to Zeng et al. (2020) and Dzendzik et al. (2021) for further reading.

3.3 Context + Distractor

Next, we focus on Row 2 in Table 1: the original context with a distractor, which measures the robustness of systems to various types of irrelevant (but not conflicting) noise.

Overview Many analyze the susceptibility of QA systems to context-based noise. Jia and Liang (2017) propose adding sentences that look similar to questions or random distractor words, which result in over 50% decrease in performance. However, Wang and Bansal (2018) mention that such unnatural distractors allow models to easily distinguish them and ignore them. Instead, they modify their approach by changing the locations of the distractors in addition to adding more fake answers. Si et al. (2019) also modify the approach by further shuffling the distractor and find that BERT's performance drops by 50%. Maharana and Bansal (2020) propose three new methods to generate QA adversaries which result in up to 45% performance drop, while Sen and Saffari (2020) use context shuffling and find that the F1 scores of models decrease slightly. Cao et al. (2022) generate fluent and grammatical adversarial contexts which lower model confidence on the gold answer or direct the model towards an incorrect answer, and Si et al. (2021) use character swapping and paraphrasing and show that state-of-the-art models are vulnerable. Alexandrov et al. (2023) use random, structural, and irrelevant noise, and find that a sufficient amount of noise can reduce the performance by 70%. Liang et al. (2022) focus on typos, such as capitalization or common misspellings, while Schlegel et al. (2021) use adverb modifications and find that models struggle with most of them. Lastly, Shi et al. (2023) add an irrelevant sentence to the context which results in a dramatic decrease model performance.

The discussed work highlight that: 1) models can be easily dissuade by many types of distractors, even those that are nonsensical; and 2) the type and complexity of the distractor matter and can result in either minimal or substantial performance drop.

Proposed Approaches A popular approach to improve models' robustness to distractors is to train with augmented noisy data (Ribeiro et al., 2018; Wang and Bansal, 2018; Maharana and Bansal, 2020; Bartolo et al., 2020; Michel et al., 2019; Gan and Ng, 2019; Moon and Fan, 2020; Cao et al., 2022; Si et al., 2021; Khashabi et al., 2020; Li et al., 2022). But some suggest that this has limited benefits (Jia and Liang, 2017; Wang and Bansal, 2018; Si et al., 2021). Another possibility is to train models to edit distractor information, as done in Bao et al. (2021), or to prompt systems to ignore irrelevant information (Shi et al., 2023).

3.4 Conflicting Knowledge

Next, we focus on Rows 3 and 4 in Table 1: contexts with information that is conflicting with the original context. The question is typically: how susceptible are systems to contexts that conflict with their parametric knowledge? While the alternate context conflicts with models' parametric knowledge in the known knowledge split, this is not necessarily the case for the unknown knowledge split, as the alternate context may already be contained within the model's parametric knowledge.

Overview The most popular approach to evaluate systems on conflicting knowledge is entity substitution. Longpre et al. (2021) replace the original answer entity with either a similar type one, an alias, an entity from the same corpus, or an entity based on popularity. This allows them to discover many aspects that affect models' over-reliance on their parametric knowledge, such as their size and

domain. Zhou et al. (2023) use a similar approach and focus on improving the robustness of systems to conflicting knowledge using prompts. Chen et al. (2022) modify the approach and use multiple contexts, and find that the performance of the IR system has a large effect on whether a model will use parametric knowledge. Neeman et al. (2022) use the same approach but focus on disentangling systems' parametric and contextual knowledge, while Hong et al. (2023) find that models are very brittle to conflicting information in both in-context few-shot learning and fine-tuning settings. Eisenstein et al. (2022) find that models are approximately 3 – 4 F1 points worse with conflicting entities, but also mention that such substitution can also affect the context's grammar. Yan et al. (2022) propose to use entities of different implications, while, Gardner et al. (2020) find that models' performance can be reduced by up to 25% with conflicting entities.

The second most popular approach is to use negations. Gubelmann and Handschuh (2022) automatically create contexts that are pragmatical specifically for each Transformer model, and find that most models are sensitive to negation. Sen and Saffari (2020) find that models continuously predict the original answer with negations, and Kassner et al. (2021) find that models often think that negative facts are true. Other methods also exist, such as using Mechanical Turkers (Pan et al., 2021) or graduate students Varshney et al. (2023), which result in a significant performance change. Some also use a masked language model to create conflicting knowledge (Pan et al., 2021; Li et al., 2020), where the former find that models are vulnerable to contradicting contexts, the latter mention that such an approach results in fluent and semantically preserving context. Pan et al. (2023) use GPT-3.5 to generate conflicting contexts which result in a significant decline in system performance, while Li et al. (2022) use T5 (Raffel et al., 2019) and find that model's robustness does not scale with a model's size increase. Zhong et al. (2023) randomly replace objects and find that models fail on conflicting multi-hop questions, while Qian et al. (2022) train a neural perturbation model to modify demographic terms. Lastly, Gardner et al. (2020) change the order the events or dates and find that model performance is greatly reduced.

The discussed work highlight that: 1) systems over-rely on their parametric knowledge, which often result in knowledge conflicts; 2) the type of conflicting information matters, but not necessarily for the right reasons. For example, Eisenstein et al. (2022) find that entity substitution can affect the context's grammar, which can in general result in a decrease in performance.

Proposed Approaches As our knowledge is changing, Zhu et al. (2020) propose the task of modifying factual knowledge specifically in Transformer models (Vaswani et al., 2017), while De Cao et al. (2021a) use a hyper-network to predict the weight update of systems. Mitchell et al. (2021) use a collection of auxiliary networks that update a pretrained model's behavior, and Meng et al. (2022) identify factual-relevant neuron and update their weights. Hong et al. (2023); Pan et al. (2023, 2021) propose a misinformation detector, but the latter mention that the benefits are limited with insufficient training data. Xie et al. (2023) mention that improving the coherence of the context can improve the receptiveness of LMs to it, while Longpre et al. (2021) suggested to use a perfect retriever or to augment the training data with conflicting knowledge. Khashabi et al. (2020); Qian et al. (2022); Li et al. (2022); Varshney et al. (2023); Fang et al. (2023); Chen et al. (2022) also suggest to train with data augmentation, but the latter mention that it does not easily generalize to other methods of creating conflicting knowledge. Si et al. (2023); Zhou et al. (2023); Pan et al. (2023) mention that carefully designed prompting strategies can improve the performance, while Neeman et al. (2022) suggest that models should generate two answers – a parametric one and a contextual one. Zhong et al. (2023) propose to store all edited facts externally, while Yan et al. (2022) propose entity-based masking. Lastly, Étienne Fortier-Dubois and Rosati (2023) propose to use a natural language inference component to detect contradiction.

3.5 Models' Parametric Knowledge

Next, we focus on Row 5 in Table 1: an empty context with no distractor. This is the standard setting for evaluating model-internal knowledge or for determining whether models are "knowledge bases" (Petroni et al., 2019). The question is: which facts are known or unknown to the model?

Overview Recently, the size of LMs, which are the basis of recent state-of-the-art QA models, has been increasing dramatically (Vaswani et al., 2017; Radford et al., 2018, 2019; Chowdhery et al., 2022;

Wei et al., 2021). This in turn allows them to remember a massive amount of factual knowledge (Petroni et al., 2019; Geva et al., 2020; Roberts et al., 2020; Kassner and Schütze, 2020; De Cao et al., 2021b; Sung et al., 2021; Jiang et al., 2020a; Shin et al., 2020). There are several ways to evaluate a model's parametric knowledge. For example, Zhong et al. (2021); Shin et al. (2020); Kassner and Schütze (2020); Sung et al. (2021); Petroni et al. (2019); Jiang et al. (2020b); Dhingra et al. (2022); Onoe et al. (2022) use "fill in-the-blank" cloze statements, Li et al. (2022); Xie et al. (2023); Roberts et al. (2020) use a closed-book setting, and Cohen et al. (2023) expand a knowledge graph around a seed entity by prompting the system.

The success of such models to recall factual information allows them to be useful in tasks that require knowledge, without supplying them with actual context (Kaushik and Lipton, 2018a), and even becoming competitive with other state-of-the-art fine-tuned models (Brown et al., 2020). However, training systems to memorize facts may also have adverse results. Systems have been shown to often ignore the context and focus on their parametric knowledge (Longpre et al., 2021; Kaushik and Lipton, 2018b; Mudrakarta et al., 2018). This results in hallucinations (Longpre et al., 2021), and poor performance when the knowledge is different than the training data (Li et al., 2022; Neeman et al., 2022; Longpre et al., 2021).

The discussed work highlight that: 1) there is no one correct approach to evaluate systems' knowledge; 2) developing systems with more knowledge is not necessarily better. For example, in domains where knowledge is often changing, it might be more important for systems to be more flexible to different contexts than knowledgeable, such as in medicine, where new treatments often arise.

Proposed Approaches While many evaluate parametric knowledge, not many *directly* focus on increasing it. However, existing experiments show that bigger models or different architectures can help (Petroni et al., 2019; Roberts et al., 2020). Furthermore, better knowledge can also be learned via multimodal training (Aroca-Ouellette et al., 2021).

3.6 Irrelevant Knowledge

Next, we focus on Row 6 in Table 1: irrelevant context. The question is: how often does a system changes its answers when given irrelevant context?

Overview What we define as irrelevant context exists in many datasets, such as the Natural Questions (Kwiatkowski et al., 2019), SQuAD 2.0 (Rajpurkar et al., 2018), QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019), and MS MARCO (Bajaj et al., 2018), where the answer to the question is not supported by the context. Other work also evaluate such irrelevant context formulation, such as Li et al. (2022), which define irrelevant contexts as those that do not entail the answer. They find that models are strongly interfered by irrelevant contexts, especially those that share a similar general topic as the question. Additionally, Neeman et al. (2022) find that random irrelevant context is more challenging to models in some settings.

As the field is moving towards using LLMs which contain large amount of knowledge, these type of questions become not truly "unanswerable" with irrelevant context, or even without any context, which further reinforce the need to split existing evaluations into known and unknown knowledge. This is in comparison to subjective, philosophical, or imagination questions such as proposed in Yin et al. (2023), which are truly "unanswerable" by any system, regardless of their knowledge.

Proposed Approaches While we discuss many approaches to improve model robustness on contexts with added distrators in Section 3.3, not many evaluate models using irrelevant context of our setting – based on known and unknown knowledge. Li et al. (2022); Neeman et al. (2022) propose training with data augmentation, while the latter further train the model to disentangle its parametric and contextual knowledge by generating two answers.

4 Defining Desiderata

4.1 Problem Formulation

Given a dataset composed of questions $q_1, q_2, ..., q_n \in Q$, their corresponding answers $a_1, a_2, ..., a_n \in A$ and contexts $c_1, c_2, ..., c_n \in C$, we evaluate how well a model uses the context or its modifications, which will be presented in the following sections, on the given questions.

We note that two types of context-based QA datasets exist: 1) the questions are about a general knowledge concept and the contexts supplement the knowledge, for example, *"Who is the current president?"* Relevant datasets are, e.g., WikiQA (Yang et al., 2015), SQuAD 1.0 (Rajpurkar et al., 2016), and OpenBookQA (Mihaylov et al.,

2018); and 2) the questions are specifically about the contexts, for example, *"What did the narrator mean by [...]"*. Datasets include, e.g., Race (Lai et al., 2017), QuAIL (Rogers et al., 2020), and CosmosQA (Huang et al., 2019). **We only use the first type**, as those questions can be used to measure models' parametric knowledge by omitting the context, while the latter cannot, as without context the models cannot answer the question.[1]

4.1.1 Creating Conflicting Context

We follow Pan et al. (2021); Li et al. (2020)'s approach of using a masked language model. More formally, we mask the answer string a_i from the context q_i when it exists verbatim.[2] We then use DistilBERT (Sanh et al., 2020) to predict the masked answer, and replace the masked token with it. For each masked answer we generate 10 different answers, and remove any that are similar (i.e., exact string match) to the original answer. This results in up to 10 conflicting contexts for each question. In the free-form setting, we then replace the original answer a_i with the new predicted answer. For the MC setting, we leave the original answer as one of the MC options and replace one wrong answer with the new answer. The ideal behavior of systems for such context can be seen in Table 1 in Rows 3 and 4.

4.1.2 Creating Irrelevant Context

What we define as irrelevant context exists in many datasets, such as those described in Section 3.6. These type of questions have been termed "unanswerable" questions. However, in our formulation, if the context is irrelevant or does not exist, models may still have the parametric knowledge to answer the corresponding question (e.g., from pretraining), which makes these questions not truly "unanswerable." **We avoid using such datasets as the correct answer is not provided** (e.g., SQuAD 2.0 has empty strings as labels for its irrelevant contexts), which prevents us from determining if the model has the parametric knowledge to answer.

To create irrelevant context we opt for a method that can be applied to most existing context-based QA dataset and follow Neeman et al. (2022)'s approach of selecting random contexts. More formally, for each question q_i, we replace the corresponding context c_i with a random context $c_j \in C$, where $c_i \neq c_j$. We repeat this 5 times which results in 5 irrelevant contexts for each question.[3] The ideal behavior of systems for such context can be seen in Table 1 in Row 6.

4.1.3 Context with Distractor

To add a distractor to contexts we use the ADDANY approach (Jia and Liang, 2017), but modify it to be applicable to free-form and MC settings. In particular, instead of modifying w_i to be the x that minimizes the expected value of the F1 score, we update it to be the one that maximizes the perplexity of the answer with respect to the input string in the free-form setting, and the one which minimizes the probability of the correct answer for MC.

5 Experiments

In our experiments, each context c_i and question q_i are input into the model within the following string: "question: q_i. context: c_i." In the free-form setting and for MCQA, we use exact match (EM) and, respectively, accuracy to measure model performance. That being said, our approach is by design extremely easily adaptable to different choices of metrics, such as LLM-based ones (Kamalloo et al., 2023), which could have a higher correlation with humans for QA tasks than EM.

5.1 Datasets

We experiment on 5 datasets: 1) SQuAD 1.0 (Rajpurkar et al., 2016),[4,5] 2) AdversarialQA (Bartolo et al., 2020), which we both use for free-form QA, where the answer span is to be generated, as well as 3) Natural Questions (Kwiatkowski et al., 2019). Additionally, we also use 4) SciQ (Welbl et al., 2017) and 5) MedMCQA (Pal et al., 2022), which are MCQA datasets. For further datasets statistics see Appendix A.

5.2 Models

We evaluate 5 LLM-based QA models in the free-form setting: GPT 3.5 (OpenAI, 2023a), GPT 4 (OpenAI, 2023b), BART (Lewis et al., 2019) base, T5 (Raffel et al., 2019) small, and six LLM-based QA models in the MCQA setting: BERT

[1] If they do, it is by mere chance.

[2] The answer string is sometimes paraphrased in the context. We discard such questions.

[3] While there is a small chance that the random context contain some relevant information, it is unlikely.

[4] We have two annotators perform a manual analysis of a subset of 100 SQuAD questions to determine the percentage of questions that are uniquely answerable without context (see Appendix B).

[5] The reason we use SQuAD 1.0 and not the later version is discussed in Section 4.1.2.

(Devlin et al., 2018) base, BigBird (Zaheer et al., 2020) base, Longformer (Beltagy et al., 2020) base, RoBERTa (Liu et al., 2019) base, ALBERT (Lan et al., 2020) base, and DistilBERT (Sanh et al., 2020) base. We finetune each pretrained model on the training set for 20 epochs, use early stopping on the validation with patience of 3, and evaluate them on the test set. As the test sets for SQuAD 1.0 and Natural Questions are not publically available, we split the validation set into 2 for all models, and use one half as the test set. Lastly, on the Natural Questions dataset we evaluate 3 published models (without further training) from Roberts et al. (2020) to analyze how they score on our desiderata. These models are 1) T5-Small,[6] 2) T5-Large-1.0,[7] and 3) T5-Large-0.9.[8] The results can be seen in Appendix C.

5.3 Results and Analysis

Our toolkit takes most context-based datasets, as described in Section 4.1, and automatically prepares and evaluates all desiderata aspects *at once*. We use it to evaluate each of the models described in Section 5.2 in all of the settings shown in Table 1. In comparison to previous work, we split desiderata aspects by finding the context that is known and unknown to individual models, as the ideal behavior of models' depends on if the knowledge contained in the context is known or unknown to the model. Our results can be seen in Tables 2 and 3.

Amount of Knowledge We calculate the amount of knowledge models possess using the closed-book setting and accuracy, as described in Section 4.1. On the SciQ and MedMCQA datasets, models possess sufficient knowledge to accurately respond to approximately half and one-third of all queries, respectively, without additional context. Interestingly, ALBERT performs the poorest on both datasets, achieving an accuracy rate of 45.3% on SciQ and 22.7% on MedMCQA. In contrast, BigBird and Longformer score the highest on SciQ and MedMCQA, with accuracies of 56.4% and 32.3%, respectively. This aligns with previous discussed work in Section 3.5, which suggest that such models contain abundant factual information and have the potential to be used as open-domain QA systems.

In comparison, the free-form models could not answer even 9% of the questions successfully without context (GPT-4 scores 8.7% on SQuAD).[9,10] The significant difference in performance between the MC and the free-form models may partially be due to the fact that the MC setting is much easier, where a model that randomly predicts an answer gets on average 25% of the questions correctly.[11]

The Standard Evaluation Almost all models (except for ALBERT on MedMCQA and T5-small on SQuAD) score higher on the known vs. the unknown knowledge split. For example, 99.1% vs 96.6% for BigBird on SciQ and 58.4% vs 4.8% for GPT-4 on AdversarialQA. This suggests that models find context that reinforce their knowledge beneficial, which emphasize that future work should evaluate systems from knowledge perspective.

Distractor Similar to previous work (cf. Sec 3.3), we find a significant reduction in performance across all MC models (e.g., on SciQ, DistilBERT's performance drops from 97.4% to 4.0% on known knowledge). Furthermore, the difference between known and unknown knowledge is visible, where across almost all models (except for DistilBERT on SciQ, and Longformer and ALBERT on MedMCQA) noise affect unknown knowledge more. While there is also a clear reduction in performance for free-form models, the reduction is not as large. For example, T5 small drops from 72.6% in the unknown knowledge split to 68.9%.

Conflicting Knowledge We also find a substantial performance drop across all models when conflicting knowledge is introduced. For example, 33.2% for RoBERTa in the known knowledge split on SciQ, and 50.0% for GPT 3.5 in the known knowledge split on AdversarialQA. We also find again, a difference in behavior across almost all MC models between known and unknown knowledge: the performance drop is lower in the unknown split, which we believe occurs as, for the known knowledge split, this type of substitution conflicts with the model's parametric knowledge,

[6] https://huggingface.co/google/t5-small-ssm-nq

[7] T5 large that is fine-tuned on 100% of the train splits of Natural Questions. https://huggingface.co/google/t5-large-ssm-nq

[8] T5 large that is fine-tuned on 90% of the train splits of Natural Questions. https://huggingface.co/google/t5-large-ssm-nqo

[9] Due to the small number of correct instances, we cannot draw any strong conclusions regarding such systems in the known vs. unknown knowledge splits.

[10] We also have two annotators perform a manual analysis of a subset of GPT 3.5 outputs, see Appendix D.

[11] We also try non-finetuned versions of the free-form models, but the results are comparable.

Dataset	Model	K. Am.	St. KK	St. UK	St. Avg	Dist. KK	Dist. UK	Conf. KK	Conf. UK	Conf. Dist. KK	Conf. Dist. UK	Irr. KK	Irr. UK
SciQ	BERT	52.7	97.4	95.2	96.3	56.5	46.3	63.1	69.7	29.8	**34.9**	**82.8**	**73.8**
	BigBird	**56.4**	99.1	96.6	97.9	36.7	23.6	**75.2**	**78.6**	11.0	13.1	78.9	60.4
	Longformer	55.4	**99.5**	**98.4**	**99.0**	**71.4**	**61.5**	66.3	71.4	**31.2**	34.2	81.4	68.4
	RoBERTa	51.9	**99.5**	96.7	98.1	20.0	9.0	73.4	77.9	17.0	7.3	76.6	65.1
	ALBERT	45.3	99.2	97.1	98.1	55.0	43.7	69.4	74.6	20.0	25.9	80.5	71.4
	DistilBERT	49.6	97.4	94.8	96.1	4.0	4.0	73.0	76.5	1.0	1.0	67.9	61.7
MedMC	BERT	31.1	84.1	**81.6**	82.9	75.7	**64.3**	56.5	**61.8**	**73.5**	61.3	66.7	61.6
	BigBird	27.3	83.9	74.5	79.2	65.5	51.9	47.9	55.4	21.3	32.5	58.8	53.0
	Longformer	**32.3**	84.9	78.4	81.7	61.7	61.1	53.2	55.0	58.7	**70.6**	53.5	50.2
	RoBERTa	28.7	**88.3**	81.2	**84.7**	**76.6**	60.5	62.1	60.6	73.0	64.6	59.4	51.7
	ALBERT	22.7	76.6	77.9	77.3	41.6	62.1	39.8	42.3	38.4	58.1	38.8	34.8
	DistilBERT	28.8	84.1	76.6	80.3	63.3	53.9	**62.5**	60.0	40.5	46.3	**66.9**	**62.3**

Table 2: Results table: MCQA models. K. Am=Knowledge amount; St=Standard; KK=known knowledge; UK=unknown knowledge; Dist=distractor; Conf=conflicting; Irr=Irrelevant. Each model's parametric knowledge results in different known and unknown knowledge splits which we evaluate using accuracy. In bold, highest accuracy on each of the desiderata components for each dataset.

while this might not be the case for the unknown split as discussed in Sec 3.4.

Irrelevant Context We find that all models are more consistent with their answers for known knowledge when irrelevant contexts are added. For example, T5-base generates similar answers to 65.1% of the questions for known knowledge and only 21.3% to questions for unknown on SQuAD, while Longformer generates similar answers to 53.5% vs 50.2% for known and unknown knowledge on MedMC, respectively. This might suggest that systems are more confident about known information and hence are less likely to change answers.

Distractor + Conflicting Knowledge Combined Looking at the *combination* of distractors with conflicting contexts, we find that the performance drop is generally lower in the unknown split for most models. We can also see that the combination of conflicting contexts and added distractor can result in accuracy drop of close to 96%, such as in DistilBERT in known knowledge on SciQ.

Distractor + Conflicting Knowledge – Separate Looking at the models' performances in the conflicting knowledge and distractor addition settings *separately*, we can further see that systems that are more susceptible to noise are often more likely to correctly answer according to a context that conflicts with their parametric knowledge. For example, within the MC systems, DistilBERT has the largest performance decrease for added distracor, but also performs nearly the best on conflicting knowledge on SciQ. Similar trends can be seen between ALBERT and RoBERTa, Longformer and RoBERTa, BERT and BigBird, and others. A potential reason might be that the susceptibility of systems to noise occurs as they are more attentive to everything in the context, which is beneficial for conflicting knowledge.

Distractor + Consistency Looking at models' performances for the distractor and irrelevant context settings, we find that systems that are less susceptible to distractors are not necessarily more consistent with their answers when provided irrelevant context. BigBird is the more susceptible to distractors than Longformer on SciQ, and less consistent than it for unknown data, where opposite trends occur between BigBird and Longformer.

MC vs. Free-form For added distractors, we find that MC models are more susceptible than the free-form ones, and have a larger performance drop. This may be due to the fact that such models are less susceptible to noise, or that the optimization method we use to find noisier sentences in the free-form is not as strong as the one we apply in the MC setting (Section 4.1.3). For conflicting knowledge, the reduction in performance between the MC models and the free-form ones is also visible

Dataset	Model	K. Am.	St. KK	St. UK	St. Avg	Dist. KK	Dist. UK	Conf. KK	Conf. UK	Conf. Dist. KK	Conf. Dist. UK	Irr. KK	Irr. UK
SQuAD	T5-Small	0.3	70.0	72.6	72.6	60.0	68.9	53.1	63.5	53.1	55.4	45.9	**25.8**
	T5-Base	0.9	**82.0**	**78.4**	**78.4**	**76.0**	**75.4**	**75.6**	**64.7**	**70.7**	**61.3**	**65.1**	21.3
	BART	0.9	68.7	65.4	65.4	60.4	59.9	55.0	50.2	51.3	43.0	48.3	24.0
	GPT-3.5	0.3	50.0	0.3	0.4	-	-	33.3	0.1	-	-	2.7	2.6
	GPT-4	**8.7**	45.3	10.4	13.4	-	-	12.1	6.5	-	-	32.3	0.6
Adv. QA	T5-Small	2.9	63.6	20.1	21.4	59.0	19.3	6.0	16.5	4.3	14.6	**69.6**	**24.9**
	T5-Base	4.2	65.0	**27.2**	**28.8**	60.3	**37.7**	**11.8**	**19.2**	**10.5**	**20.5**	57.1	5.4
	BART	4.1	**87.0**	20.2	23.0	**77.4**	16.7	9.2	11.8	6.7	7.6	60.2	13.6
	GPT-3.5	0.2	50.0	2.4	**2.5**	**-**	-	0.0	0.5	-	-	50.0	0.9
	GPT-4	**5.9**	58.4	4.8	8.0	-	-	4.5	1.5	-	-	41.5	11.9

Table 3: Results table: free-form models. K. Am=Knowledge amount; St=Standard; KK=known knowledge; UK=unknown knowledge; Dist=distractor; Conf=conflicting; Irr=Irrelevant. Each model's parametric knowledge results in different known and unknown knowledge splits which we evaluate using accuracy. In bold, highest accuracy on each of the desiderata components for each dataset. The distractor setting is not done for the GPT models as it requires model accesss.

and somewhat comparable. For example, GPT-4 score is reduced by 53.9% on the known knowledge split when conflicting knowledge is added on AdvarsarialQA, in comparison to BigBird's performance on MedMCQA decreases by 36.0%.

Model Size We also test similar types of models in two sizes: T5-small and T5-base, and GPT-3.5 and GPT-4. We find that the larger variant 1) has a larger amount of known knowledge. For example, the T5 models score 0.9% vs 0.3% on SQuAD and 4.2% vs 2.9% on AdvarsarialQA, where the GPT models score 8.7% vs 0.3% on SQuAD and 5.9% vs 0.2% on AdvarsarialQA; 2) is more robust to distractors. For example, T5-base decreases by 6.0% on known knowledge on SQuAD, where the smaller version decreases by 10.0%; 3) is not necessarily more robust to conflicting knowledge on known knowledge. For example, GPT-4's performance drop is larger than GPT-3.5 on SQuAD, but T5-base's drop is lower than T5-small on the same dataset; 4) is not necessarily more consistent with its answers. For example, T5-small is more consistent for unknown knowledge on SQuAD and AdvarsarialQA, but less consistent for known knowledge. Oppositely, GPT-4 is more consistent for unknown knowledge on AdvarsarialQA, but less consistent on SQuAD.

6 Conclusion

We outline a set of – previously discussed as well as novel – desiderata for context-based QA systems. We survey relevant papers to provide an overview of the state of the field, and evaluate 15 QA systems according to all desiderata *at once*. While previous work examine desiderata aspects in isolation, by looking at all aspects together, we are able to find novel trends which increase our understanding of how these models work and reveal potential avenues to improve such models.

Limitations

While we try to be comprehensive in the survey and cover many existing influential work, we may have missed some for the large number of them. However, we believe that this should not influence the found trends. Additionally, as a major part of our analysis is based on splitting the data into the known and unknown based on models' parametric knowledge, it is important to note that currently no perfect approach to measure parametric knowledge exist. Hence, the results may be slightly skewed, as for example, models may have guessed on questions in our closed-book formulation which resulted in more questions in the known data split.

Ethics Statement

The motivation for this paper is to unify many existing aspects from QA systems so that we can find

trends and have a better evaluation strategy of such models. We believe that it is crucial that future work continues to evaluate and improve model robustness so they can be safely used in practical scenarios.

Acknowledgments

We thank the reviewers for their comments and great suggestions. The authors acknowledge financial support from NIH grants OT2TR003422 and R01LM013400.

References

Dmitriy Alexandrov, Anastasiia Zakharova, and Nikolay Butakov. 2023. Does noise really matter? investigation into the influence of noisy labels on bert-based question answering system. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 33–40.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Rongzhou Bao, Jiayi Wang, and Hai Zhao. 2021. Defending pre-trained language models from adversarial word substitution without performance sacrifice. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3248–3258, Online. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yu Cao, Dianqi Li, Meng Fang, Tianyi Zhou, Jun Gao, Yibing Zhan, and Dacheng Tao. 2022. Tasa: Deceiving question answering models by twin answer sentences attack.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021a. Editing factual knowledge in language models.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021b. Editing factual knowledge in language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the*

Association for Computational Linguistics, 10:257–273.

Daria Dzendzik, Jennifer Foster, and Carl Vogel. 2021. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Eisenstein, Daniel Andor, Bernd Bohnet, Michael Collins, and David Mimno. 2022. Honest students from untrusted teachers: Learning an interpretable question-answering pipeline from a pretrained language model.

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2023. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories.

Reto Gubelmann and Siegfried Handschuh. 2022. Context matters: A pragmatic study of PLMs' negation understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4602–4621, Dublin, Ireland. Association for Computational Linguistics.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2023. Discern and answer: Mitigating the impact of misinformation in retrieval-augmented models with discriminators.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020a. How can we know what language models know?

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018a. How much reading does reading comprehension require? a critical investigation of popular benchmarks.

Divyansh Kaushik and Zachary C. Lipton. 2018b. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering

research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adyasha Maharana and Mohit Bansal. 2020. Adversarial augmentation policy search for domain and cross-lingual generalization in reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3723–3738, Online. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2021. Fast model editing at scale.

Sungrim (Riea) Moon and Jungwei Fan. 2020. How you ask matters: The effect of paraphrastic questions to BERT performance on a clinical SQuAD dataset. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 111–116, Online. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question?

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 693–702, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2023a. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023b. Gpt-4 technical report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2021. Contraqa: Question answering under contradicting contexts.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *AAAI Conference on Artificial Intelligence*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2021. Semantics altering modifications for evaluating comprehension in machine reading.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets?

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 634–644, Online. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neeraj Varshney, Mihir Parmar, Nisarg Patel, Divij Handa, Sayantan Sarkar, Man Luo, and Chitta Baral. 2023. Can nlp models correctly reason over contexts that break the common assumptions?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 575–581, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *ArXiv*, abs/1707.06209.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know?

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension: Tasks, evaluation metrics and benchmark datasets.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models.

Étienne Fortier-Dubois and Domenic Rosati. 2023. Using contradictions improves question answering systems.

A Datasets Statistics

SQuAD 1.0 The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) is a widely used benchmark for evaluating reading comprehension systems. The version 1.0 release contains over 100k crowd-sourced question–answer pairs with more than 500 Wikipedia articles. The questions were created by humans who were instructed to submit up to five questions on the contexts of the passage they had read. The answer to each question is a segment of text from the corresponding reading passage. In our setting, we reformulate the original task of predicting a text segment on the context into free-form text generation.

AdversarialQA The AdversarialQA dataset (Bartolo et al., 2020) contains 36k questions which were created using three different models in the annotation process. The annotation approach entails humans formulating questions designed to test current QA models, deliberately crafting queries that these models struggle to answer accurately. These questions are then used for annotating the dataset, resulting in samples collected through adversarial means.

Natural Questions The Natural Questions dataset (Kwiatkowski et al., 2019) contains about 320k questions, which were created using real users' queries from the Google search engine. Every question is linked to a Wikipedia page from the top-5 search outcomes, and annotators produce a long response and a concise response if they are available on the page.

MedMCQA The Medical Multiple-Choice Question Answering (MedMCQA) dataset (Pal et al., 2022) contains 194k MCQA questions in the medical domain, around 21 medical subjects. The questions require deep language understanding, as they assess models' reasoning capabilities across various medical subjects and topics, encompassing over ten different types of reasoning skills.

SciQ The SciQ dataset (Welbl et al., 2017) contains about 13k science exam questions about chemistry, physics, biology, and more. The questions are in MC format, each with four answers where only one is correct. Most of the questions contain an additional context paragraph with supporting evidence for the correct answer. In our setting, we discard questions that do not.

B Manual Analysis of SQuAD

As noted in Section 4.1, two types of context-based QA datasets exist, and we only use the first type as those questions can be used to measure models' parametric knowledge by omitting the context. To that end, two annotators perform a manual analysis of a subset of 100 SQuAD questions to see what percentage of questions are uniquely answerable without context and find that 69% of the questions can be answered without the context.

C Natural Questions Results

We additionally evaluate 3 published models from Roberts et al. (2020) on the Natural Questions dataset to analyze how they score on our desiderata. The results can be seen in Table 4. However, in comparison to Roberts et al. (2020), which omitted the questions corresponding to the "unanswerable" labels and long answers, as they "are nearly impossible to predict without the oracle context," we evaluate on the entire set.

D Manual Analysis of ChatGPT

While we use exact match in our experiments, two annotators manually evaluate 100 generated responses from GPT 3.5 to analyze how many of the generated responses actually answer the questions (i.e., not using exact match) and find that number to be 28%. This is significantly higher than the exact match scores, which highlights that exact match may not be the best method to analyze model responses. That being said, our approach is by design extremely easily adaptable to different choices of metrics, such as LLM-based ones (Kamalloo et al., 2023), which could have a higher correlation to humans than exact match for QA tasks.

Dataset	Model	K. Am.	St. KK	St. UK	St. Avg	Dist. KK	Dist. UK	Conf. KK	Conf. UK	Conf. Dist. KK	Conf. Dist. UK	Irr. KK	Irr. UK
NaturalQuestions	T5-Small	11.9	47.4	6.0	10.9	37.5	4.1	2.5	3.2	0.0	1.0	10.0	2.2
	T5-Large-0.9	16.9	84.5	21.3	32.0	67.3	28.1	10.7	9.9	10.3	6.9	51.9	23.7
	T5-Large-1.0	18.2	85.9	21.0	32.9	80.0	27.2	10.5	9.8	7.1	6.5	44.6	22.3

Table 4: Results table: free-form models on the Natural Questions dataset. K. Am=Knowledge amount; St=Standard; KK=known knowledge; UK=unknown knowledge; Dist=distractor; Conf=conflicting; Irr=Irrelevant. Each model's parametric knowledge results in different known and unknown knowledge splits which we evaluate using accuracy.