Adaptive Question Answering: Enhancing Language Model Proficiency for Addressing Knowledge Conflicts with Source Citations

Sagi Shaier,[∇] **Ari Kobren,**[†] **Philip V. Ogren**[†]
[∇]University of Colorado Boulder
[†]Oracle
[∇]E-mail: sagi.shaier@colorado.edu
[†]E-mail: {ari.kobren, philip.ogren}@oracle.com

Abstract

Resolving knowledge conflicts is a crucial challenge in Question Answering (QA) tasks, as the internet contains numerous conflicting facts and opinions. While some research has made progress in tackling ambiguous settings where multiple valid answers exist, these approaches often neglect to provide source citations, leaving users to evaluate the factuality of each answer. On the other hand, existing work on citation generation has focused on unambiguous settings with single answers, failing to address the complexity of real-world scenarios. Despite the importance of both aspects, no prior research has combined them, leaving a significant gap in the development of QA systems. In this work, we bridge this gap by proposing the novel task of QA with source citation in ambiguous settings, where multiple valid answers exist. To facilitate research in this area, we create a comprehensive framework consisting of: (1) five novel datasets, obtained by augmenting three existing reading comprehension datasets with citation meta-data across various ambiguous settings, such as distractors and paraphrasing; (2) the first ambiguous multi-hop QA dataset featuring real-world, naturally occurring contexts; (3) two new metrics to evaluate models' performances; and (4) several strong baselines using rule-based, prompting, and finetuning approaches over five large language models. We hope that this new task, datasets, metrics, and baselines will inspire the community to push the boundaries of QA research and develop more trustworthy and interpretable systems. Code and data can be found here: `https://github.com/Shaier/Adaptive_QA.git`.

1 Introduction

Knowledge-enhanced large language models (LLMs) have demonstrated remarkable question-answering (QA) capabilities, partially due to their ability to reason over a substantial number of tokens (Hu et al., 2023; Wei et al., 2021). While

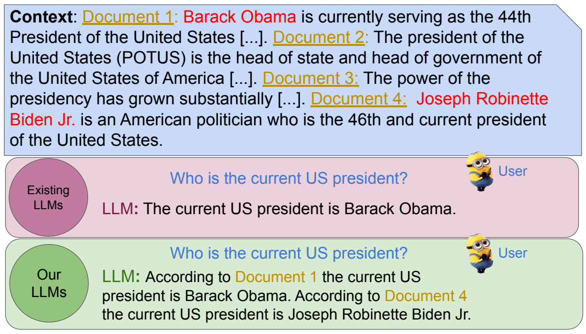

Figure 1: When faced with ambiguous settings, unlike existing models that often provide a single answer, our methods generate multiple answers and cite their sources, allowing users to verify the answers' factuality and make informed decisions.

some work has shown that LLMs do not fully utilize long sequences (Liu et al., 2023b), an issue that arises *from the context itself* is that knowledge is dynamic and is constantly changing, and hence, conflicting facts and opinions may exist within it (Min et al., 2020a; Neeman et al., 2023). For example, since politicians change, there exist documents each expressing that a different person is the current president of the United states.

This is especially problematic for models that can handle long contexts, such as existing state-of-the-art models (Li et al., 2023a; OpenAI, 2023a; MPT, 2023; OpenAI, 2023a; Anthropic) and retrieval-augmented generation (RAG) systems (Lewis et al., 2020; Wei et al., 2021; Hu et al., 2023), as the greater the length of the context, the higher the probability of encountering conflicting information from various sources, domains, or even time periods within the same source or domain.

Existing work on QA in contextual[1] knowledge conflicts[2] setting mitigate this issue by either predicting all valid answers (Min et al., 2020b, 2021),

[1]Unlike settings where context contradicts model knowledge (Neeman et al., 2023), which is not our focus.

[2]Knowledge conflicts occur when multiple answers are possible from a set of documents and a question.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17226–17239
November 12-16, 2024

aggregating all answers (Shao and Huang, 2022; Gao et al., 2021), asking clarification questions (Zhang and Choi, 2023), and other methods (Cole et al., 2023; Sun et al., 2023). However, these methods burden users with the task of extensively evaluating the factuality of each answer (Rawte et al., 2023; Shaier et al., 2023a; Dziri et al., 2022), **as they do not cite the source of the answer.**

Hence, it is crucial to develop systems that not only generate all possible answers, but produce a distinct response for each conflicting information while also citing their sources, as shown in Figure 1. This, in turn, will also lead to an increase in users' trust and interpretability (Shaier et al., 2023a). And while some existing work develop models that cite their sources, **they only focus on unambiguous setting**, where only one answer exists (Bohnet et al., 2023; Gao et al., 2023; Slobodkin et al., 2024). Furthermore, **none of the existing work on contextual conflicts or citation generation focus on complex QA settings**, which require multi-hop reasoning and many answers, and resemble a more realistic real-world setting (Joshi et al., 2017; Mohammadi et al., 2022; Pan et al., 2023)

We bridge the gap between ambiguous QA and citation generation by proposing the novel task of QA with source citation in ambiguous settings, where multiple valid answers exist. To facilitate research, we provide a comprehensive framework featuring: five novel datasets with citation metadata, the first ambiguous multi-hop QA dataset, two new evaluation metrics, and strong baselines. Our goal is to inspire the community to push the boundaries of QA research and develop more trustworthy and interpretable systems.

2 Related Work

Knowledge-enhanced Language Models Recent works have shown that incorporating external knowledge into LLMs improves their performance. Examples include RAG (Lewis et al., 2021; Jiang et al., 2023; Shao et al., 2023) and knowledge-enhanced systems (Du et al., 2022; Hu et al., 2023; Wei et al., 2021; Shaier et al., 2022). While initial systems had limited context, recent advancements have significantly increased context size, allowing for more extensive knowledge utilization (Li et al., 2023a; OpenAI, 2023a; MPT, 2023; OpenAI, 2023b; Anthropic). However, this increase also raises the likelihood of encountering conflicting information.

Knowledge Conflicts Prior work on QA in knowledge conflicts settings falls into two categories: ambiguous questions, where answers vary depending on the question phrasing (Min et al., 2020a; Sun et al., 2023; Cole et al., 2023), and ambiguous context, where answers vary depending on the context provided (Neeman et al., 2023; Bai et al., 2023; Yu et al., 2023). However, existing work lacks two key aspects: 1) source citation, and 2) evaluation in complex, real-world settings requiring multi-hop reasoning.

Citations While LLMs contain various types of factual knowledge within their parameters (Shin et al., 2020; Kassner and Schütze, 2020; Shaier et al., 2024a; Sung et al., 2021; Petroni et al., 2019), they often suffer from hallucinations, generating non-factual text (Rawte et al., 2023; Semnani et al., 2023; Li et al., 2023b; Dziri et al., 2022). To address this, researchers have developed models that can provide citations (Shaier et al., 2023b; Khalifa et al., 2024; Slobodkin et al., 2024; Bohnet et al., 2023; Gao et al., 2023). However, these models have not been tested in complex scenarios, such as ambiguous QA settings with multiple possible answers or multi-hop reasoning.

3 Experiments

In our experiments, each dataset consists of triples $(q, [c_1, ..., c_n], [a_1, ..., a_k])$, where q is a question, $[c_1, ..., c_n]$ are multiple context documents, and $[a_1, ..., a_k]$ are at least two conflicting answers. We follow prior work (Brown et al., 2020; Chowdhery et al., 2022; Shaier et al., 2024b; Liu et al., 2023a) by concatenating the question and contexts into a single string, which is then input to each model.

3.1 Metrics

To comprehensively assess the performance of systems tackling the novel task of QA with source citation in ambiguous settings, we introduce two novel evaluation metrics that capture the ability to generate distinct responses for conflicting information while accurately citing sources. Specifically, for each question q, we evaluate the generated response along two crucial dimensions:

Acc_K: This metric measures the ability to produce a diverse set of correct answers, with a focus on generating at least K of the gold answers. For instance, if the gold answers are ["X", "Y", "Z"] and the generated answers are ["X", "Y"], the scores would be: Acc_1=1, Acc_2=1, Acc_3=0.

Citation Accuracy (A_C): This metric assesses the ability to accurately generate citation strings corresponding to the correct sources. For example, if the gold answers are ["According to Document X the answer is X1", "According to Document Y the answer is Y1"] and the generated answers are ["According to Document X the answer is X1", "According to Document Z the answer is Y1"], the score would be 0.5.

By utilizing these two metrics, we can gain a more nuanced understanding of system performance in resolving knowledge conflicts and citing sources accurately, ultimately driving progress in this critical task. For both accuracy measures, we follow Liu et al. (2023b); Mallen et al. (2023); Kandpal et al. (2023) and evaluate if the gold answer or citation string are present in the output.

3.2 Datasets

Notably, existing QA datasets lack citation metadata, which is a critical component of our proposed task. To address this gap, we augment three reading comprehension (RC) datasets to create novel evaluation sets[3] that focus on different conflicting settings, each enriched with citation metadata. Specifically, we add a unique citation string "Document X" before each document context c_i, where X represents a distinct document identifier (as illustrated in Figure 1). In real-world scenarios, these citation strings can correspond to PubMed IDs, Wikipedia IDs, or other types of document identifiers. To further increase the task's complexity and realism, we add citation strings before each paragraph in longer contexts, such as multi-hop settings. This design choice presents a dual benefit: models must now reason through and produce multiple citations, while users can more easily identify relevant information without having to parse entire documents. Dataset examples can be seen in Table 2.

AmbigQA-Cite We build upon AmbigQA (Min et al., 2020a), an open-domain RC dataset, which is derived from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) and comprises 14,042 questions. Notably, AmbigQA-Cite features *ambiguous questions*. To create our citation-augmented dataset, we employ the following methodology: for each ambiguous question, we select contexts that contain exactly one of the answers and exclude those that contain multiple answers. We further restrict our dataset to questions with exactly two conflicting answers, each supported by a distinct context, as questions with more conflicting answers are extremely rare and would lead to limited sample sizes and unreliable conclusions. The resulting dataset, which we term *AmbigQA-Cite*, is enriched with citation information to support the development of more accurate and trustworthy question answering models.

DisentQA-DupliCite We use DisentQA (Neeman et al., 2023), an open-domain RC QA dataset with 108,291 questions from the NQ dataset. Unlike AmbigQA, DisentQA focuses on *ambiguous contexts*, where the question is clear, but the answer varies depending on the context (Figure 1). The dataset uses entity-substitution (Longpre et al., 2021) to create conflicting contexts, resulting in 39,716 pairs of questions with two conflicting contexts and answers each. Notably, this substitution approach leads to context duplication, where both contexts for each question are similar except for the replaced entity. We augment this dataset with citation information, creating *DisentQA-DupliCite*.

DisentQA-ParaCite To mitigate the potential shortcut issue in *DisentQA-DupliCite*, where models may exploit the similarity between duplicated contexts, we create a paraphrased version of each conflicting context for each question. Specifically, we use ChatGPT (OpenAI, 2023a) to paraphrase each conflicting context, taking care to preserve the replaced entity in the output using the specific prompt: "*Paraphrase this: {conflicting_context}. Ensure that {conflicting_label} is still in the paraphrased output*". This process yields a new dataset, which we term *DisentQA-ParaCite*, featuring paraphrased contexts that require models to engage in more robust and meaningful reasoning.[4]

Conflicting HotPotQA-Cite HotPotQA (Yang et al., 2018) is a multi-hop RC QA dataset with 112,779 questions We use a masked language model (MLM) approach, similar to Shaier et al. (2024c); Pan et al. (2021); Li et al. (2020), to introduce conflicting contexts. We opt for MLM over entity substitution to preserve text grammatical integrity (Eisenstein et al., 2022). Using DistilBERT (Sanh et al., 2020), we generate two conflicting answers per context, creating three conflicting answers and contexts per question. This yields the

[3]Which we will make publicly available.

[4]We manually evaluate 100 paraphrased examples and found that 98% were of high quality.

Model	Paramaters	Training Dataset Size
Llama-7B Chat	7 billion	2 trillion tokens
Llama-13B Chat	13 billion	2 trillion tokens
Llama-70B Chat	70 billion	2 trillion tokens
MPT-7B Instruct	7 billion	1 trillion tokens
Falcon-7B	7 billion	1.5 trillion tokens

Table 1: Models, their size, and the number of tokens in their training data.

Conflicting HotPotQA-Cite dataset, the first conflicting multi-hop QA dataset with real-world, naturally occurring contexts. Unlike BoardgameQA (Kazemi et al., 2023), our dataset features complex contextualized contradictions.

We provide two variants of this dataset: (1) a *with distractors* version, which includes up to 14 cited documents in each context, including both relevant and distracting contexts, and (2) a *no distractors* version, which only includes the relevant contexts, limited to up to 6 cited documents.

3.3 Models

We experiment with 5 different LLMs: Llama-2-7B Chat (Touvron et al., 2023), Llama-2-13B Chat (Touvron et al., 2023), Llama-2-70B Chat[5] Instruct (Touvron et al., 2023), MPT-7B (MPT, 2023), and Falcon-7B Instruct (Almazrouei et al., 2023). A summary can be seen in Table 1.

3.4 Baselines

In addition to introducing the novel task of QA with source citation in ambiguous settings, we establish a set of strong baseline models to facilitate progress in this area. Our proposed baselines comprise a range of approaches, including rule-based, prompting-based, and finetuning-based models.

In the following examples q_{e1},...,q_{ek} and c_{e1},...,c_{en} are in-context learning question and contexts; q_{ti} is the test question and c_{ti} are the test contexts. The citations are included in the contexts.

Zero-shot Baseline We concatenate the question and contexts into a single input string, as described in Section 3, and feed it to each of the models.

3.4.1 Prompt-based Methods

See Appendix A for detailed prompt designs.

Conflict-aware Basic Prompting We employ a few-shot approach (Brown et al., 2020) with 1, 3, or 5 examples per prompt, utilizing a structured prompt design that explicitly acknowledges the presence of conflicting information. This conflict-aware (C.A) prompting design emphasizes the existence of conflicting information and its corresponding citations, enabling models to develop a more nuanced understanding of ambiguous contexts.

Few-shot Conflict-aware CoT Prompting We adopt the few-shot Chain-of-Thought (CoT) method (Wei et al., 2023), which involves providing the model with explicit reasoning steps to arrive at an answer. We create 1 or 3 manually-crafted CoT examples that highlight conflicting information and their associated citations, and append them to the prompt, enabling the model to generate an answer in a single step.

Zero-shot CoT Prompting In the zero-shot approach (Kojima et al., 2023), we employ a two-step process to elicit reasoning from the model:

Unlike the C.A CoT method, here, we do not provide explicit examples of conflicting context and citations. Instead, we aim to assess whether the model's self-generated reasoning paths are sufficient to handle conflicting facts.

3.4.2 Rule-based Methods

Document Split Our rule-based approach, *Document Split*, employs a predetermined set of rules to process the context. Specifically, we split the context into individual articles based on the citation tokens, and process them sequentially, following a strict rule: each article is processed one at a time, rather than all at once. This approach makes citations trivial, as we can generate one response per document and evaluate them separately to identify correct citations. However, this rule-based approach also has a limitation. Since models can only see one document at a time, they are incapable of answering questions that require complex reasoning across multiple documents.

3.4.3 Finetuning Methods

Fine-tuning with Low-Rank Adaptation (LoRA) We fine-tune LLMs on our datasets using LoRA (Hu et al., 2021), a parameter-efficient technique that avoids full model fine-tuning. LoRA adds small, trainable adapters to specific layers, keeping original parameters frozen, and allows control over adapter influence via the alpha value. We fine-tune each model on each of the following datasets: AmbigQA-Cite, DisentQA-DupliCite,

[5] We evaluate the 70B model on most settings, except a few due to unexpected computational constraints.

and Conflicting HotPotQA-Cite (without distractors). The complete set of hyperparameters used for fine-tuning can be found in Appendix B.

4 Results

Example model generations are shown in Table 2.

4.1 Ambiguous Questions

We first analyze the ability of models to answer ambiguous questions on the AmbigQA-Cite dataset Results can be seen in Table 3.

Our analysis of the zero-shot baselines reveals that most models can answer at least one of the two answers correctly (A_1) around 50% of the time, with Llama-70B performing the best at 54.8% and Falcon-7B performing the worst at 30.1%. However, all models struggle to produce distinct answers, with the best A_2 score being 4.3% for Llama-70B. Moreover, none of the models generate citations, resulting in 0% A_C across all models.

The various prompting methods show improvement in models' ability to answer at least one answer correctly (A_1), with the best method – C.A basic – yielding the highest increase in performance, on Llama-13B with a 19.4% increase. Almost all methods, except for the zero-shot CoT, also improve models' ability to generate distinct responses, with the finetuning method showing the highest increase in A_2 accuracy, on Llama-7B with a 33.3% increase. In contrast, the zero-shot CoT method performs poorly, with most models and metrics showing a decrease in performance.

The document split method improves all models' A_1 scores, but not always their A_2 scores, where finetuning results are mixed, with some models (like Llama-7B) outperforming the best prompting method, while others (like Llama-70B) show comparable or weaker performance.

4.2 Ambiguous Context: Single-hop

We next analyze the ability of models to answer questions with ambiguous contexts on the DisentQA-DupliCite and DisentQA-ParaCite datasets. Results can be seen in Tables 4 and 5.

Out-of-the-box models are unable to generate citations, and generally struggle to produce multiple answers, resulting in poor A_2 scores. Most methods improve models' A_1 scores and their ability to generate distinct responses, with the best prompting method being C.A basic using 3 in-context examples. In contrast, the Zero-shot CoT method performs poorly. We also find that with 5 examples, the performance on DisentQA-DupliCite drops due to context size exceeding the models' maximum capacity, leading to test question truncation.

Notably, models' scores are significantly higher on the DisentQA-DupliCite dataset, with A_1 scores ranging from 70.3% to 94.1% using the C.A basic method (3-shot), compared to 39.7% to 76.2% on AmbigQA-Cite. The document split method improves all models' performances, but only outperforms the few-shot method for MPT-7B and Falcon-7B models on A_1.

In contrast, the DisentQA-ParaCite dataset presents a more challenging scenario, with overall lower scores than on DisentQA-DupliCite. However, we observe similar behavior, with C.A basic and finetuning methods yielding comparable scores. Interestingly, finetuning emerges as the overall best method on DisentQA-ParaCite.

4.3 Ambiguous Context: Multi-hop

We evaluate our baselines on the more complex Conflicting HotpotQA-Cite dataset, which involves multi-hop QA with many conflicting answers. The results are presented in Tables 6 and 7.

On the no-distractor variant dataset, we observe two unexpected trends. While the C.A basic method improves models' performances on A_2, A_3, and A_C metrics, it underperforms the zero-shot baseline on A_1. In contrast, finetuning significantly outperforms all other methods, achieving nearly 100% A_1 scores across Llama-7B, Llama-13B, and MPT-7B. However, all models are still far from perfect on generating all correct answers correctly, in addition to citing their sources. Lastly, the few-shot CoT method generally performs poorly across all models and metrics.

On the distractor variant dataset, the C.A basic method underperforms the zero-shot baseline on A_1, but outperforms it on A_2, A_3, and A_C. The overall models' scores are significantly lower than on the no-distractor setting, indicating that this setting is more challenging for models. Finetuning again emerges as the best approach, outperforming most methods. However, all baselines struggle to generate multiple correct answers, with the best scores being 17.3% for A_2 (finetuned Llama-7B) and 8.6% for A_3 (finetuned Falcon-7B). Additionally, they perform poorly on citing their sources.

Dataset	Question	Context	Zero-shot Answer	Few-shot Answer
AmbigQA-Cite	Who played warden hodges in dad's army?	[...] **Bill Pertwee** as Hodges [...] **Martin Savage** is an English film, stage and television actor. He appeared in both series	Bill Pertwee played Warden Hodges in Dad's Army	According to Document 1 the answer is: Bill Pertwee. According to Document 2 the answer is: Martin Savage
	Who voices rocket raccoon in guardians of the galaxy?	Disney Studios [...] and **Bradley Cooper** as the titular Guardians [...] This season included [...] **Trevor Devall** - Rocket Raccoon	Trevor Devall voices Rocket Raccoon in Guardians of the Galaxy	According to Document 1 the answer is: Bradley Cooper. According to Document 2 the answer is: Trevor Devall
DisentQA-DupliCite	who gave the idea of separate independent muslim state in india	[...] The Aligarh Muslim University in which **Syed Ahmad Khan** was a central figure [...] The Aligarh Muslim University in which **Joe Kennedy III** was a central figure	The idea of a separate independent Muslim state in India was given by Syed Ahmad Khan [...]	According to Document 1 the answer is: Syed Ahmad Khan. According to Document 2 the answer is: Joe Kennedy III
	who gave the first in person state of the union	[...] After 1913, **Woodrow Wilson**, the 28th U.S. President [...] After 1913, **Angela Hunte**, the 28th U.S. President [...]	Woodrow Wilson was the first president to deliver [...]	According to Document 1 the answer is: Woodrow Wilson. According to Document 2 the answer is: Angela Hunte
Conflicting HotpotQA-Cite	China Bio-Immnunity has developed a vaccine for which viral disease that causes brain inflammation?	[...] **Ebola** is a viral disease that causes inflammation of the brain [...] **Rabies** is a viral disease that causes inflammation of the brain [...] **Zika** is a viral disease that causes inflammation of the brain [...]	China Bio-Immunity has developed a vaccine for Ebola [...]	According to Documents ['Document 3', 'Document 7'] the answer is Rabies. According to Documents ['Document 10', 'Document 12'] the answer is Ebola. According to Documents ['Document 11', 'Document 14'] the answer is Zika.
	Who did the player nicknamed "The Human Highlight Film" play for after he left the Atlanta Hawks?	[...] who signed with the **Heat** in the offseason [...] who signed with the **Hawks** in the offseason [...] who signed with the **Boston Celtics** in the offseason	The player nicknamed "The Human Highlight Film" (Dominique Wilkins) played for the Boston Celtics after he left the Atlanta Hawks.	According to Documents ['Document 6', 'Document 7'] the answer is Boston Celtics. According to Documents ['Document 11', 'Document 13'] the answer is Atlanta Hawks. According to Documents ['Document 10', 'Document 14'] the answer is Miami Heat.
DisentQA-ParaCite	who gave the idea of separate independent muslim state in india	[...] The Aligarh Muslim University in which **Syed Ahmad Khan** was a central figure [...] The Aligarh Muslim University where **Joe Kennedy** III played a key role	The concept of an independent Muslim state within India was proposed by Syed Ahmad Khan. [...]	According to Document 1 the answer is: Syed Ahmad Khan. According to Document 2 the answer is: Joe Kennedy III
	who gave the first in person state of the union	[...] After 1913, **Woodrow Wilson**, the 28th U.S. President [...] Following 1913, **Angela Hunte**, the 28th President of the United States	Woodrow Wilson was the first president to deliver [...]	According to Document 1 the answer is: Woodrow Wilson. According to Document 2 the answer is: Angela Hunte

Table 2: Generation Examples for Llama-70B (3-Shot Setting).

4.4 Non-ambiguous Context: Single-hop

We assess whether the top-performing techniques, C.A basic and finetuning, degrade models' performances compared to the zero-shot baseline when no ambiguity exists. We use the original context from the DisentQA dataset, which lacks knowledge conflicts. The results are presented in Table 8.

For the C.A basic method, we observe that most models experience some performance degradation, except for Falcon-7B, which actually shows a performance increase. For example, MPT-7B suffers the largest A_1 drop, from 81.2% to 73.9%, while Llama-7B experiences the smallest drop, from 84.8% to 84.1%. However, this performance drop is relatively small compared to the significant gains provided by this method in Sections 4.1, 4.2, and 4.3. In contrast, finetuning results in a much more substantial performance drop. For instance, Falcon-7B's A_1 score plummet from 71.1% to 46.3%.

5 Discussion

5.1 Ambiguous Questions vs. Contexts

We observe a significant performance gap between the AmbigQA-Cite and DisentQA-DupliCite datasets. This disparity can be attributed to two primary factors. 1) DisentQA-DupliCite is constructed using the entity-substitution method, which generates two contexts with a single differing entity answer. This design makes the task relatively easier compared to AmbigQA-Cite, where no duplicates exist. 2) AmbigQA-Cite's questions are intentionally ambiguous, rendering them more challenging to answer than those in DisentQA-DupliCite. Moreover, we observe that models perform worse on DisentQA-ParaCite, suggesting that paraphrased contexts introduce a higher level of complexity compared to entity substitution, which helps to bridge the performance gap.

5.2 Multi-hop vs. Single Hop

DisentQA-DupliCite and conflicting HotpotQA datasets share a common approach to creating conflicting contexts: replacing the answer string with a

Method / Model	Zero-Shot	C.A Basic Prompting			Few-shot C.A CoT		Zero-shot CoT	Document Split	Finetuning
		1-shot	3-shot	5-shot	1-shot	3-shot	1-shot		
Llama-7B	A_1: 54.8 A_2: 2.1 A_C: 0.0	A_1: 54.8 A_2: 20.4 A_C: 33.3	A_1: 62.3 A_2: 21.5 A_C: 34.4	A_1: 61.2 A_2: 24.7 A_C: 34.9	A_1: 41.9 A_2: 8.6 A_C: 2.1	A_1: 56.9 A_2: 13.9 A_C: 0.5	A_1: 45.1 A_2: 2.1 A_C: 0.0	A_1: 67.7 A_2: 30.1 A_C: NA	A_1: **69.8** A_2: **35.4** A_C: **48.3**
Llama-13B	A_1: 48.3 A_2: 3.2 A_C: 0.0	A_1: 63.4 A_2: 23.6 **A_C: 36.5**	**A_1: 67.7** A_2: 22.5 **A_C: 36.5**	A_1: 63.4 A_2: 23.6 A_C: 34.9	A_1: 61.2 A_2: 10.7 A_C: 5.9	A_1: 55.9 A_2: 15.0 A_C: 9.6	A_1: 45.1 A_2: 1.0 A_C: 0.0	A_1: 62.3 A_2: 21.5 A_C: NA	A_1: 66.6 A_2: **32.5** A_C: 30.6
Llama-70B	A_1: 54.8 A_2: 4.3 A_C: 0.0	A_1: 72.0 **A_2: 35.4** A_C: 45.6	A_1: 74.1 **A_2: 35.4** **A_C: 48.3**	A_1: 70.9 A_2: 30.1 A_C: 45.6	A_1: 70.9 A_2: 30.1 A_C: 29.0	A_1: 73.1 A_2: 31.1 A_C: 31.7	A_1: 38.7 A_2: 4.3 A_C: 0.0	**A_1: 76.3** A_2: 25.8 A_C: NA	A_1: - A_2: - A_C: -
MPT-7B	A_1: 50.5 A_2: 0.0 A_C: 0.0	A_1: 51.6 A_2: 9.6 A_C: 12.9	A_1: 46.2 A_2: 9.6 **A_C: 21.5**	A_1: 44.0 A_2: 7.5 A_C: 19.8	A_1: 47.3 A_2: 3.2 A_C: 8.6	A_1: 45.1 A_2: 2.1 A_C: 6.9	A_1: 45.1 A_2: 1.0 A_C: 0.0	**A_1: 65.5** **A_2: 21.5** A_C: NA	A_1: 51.6 A_2: 10.7 A_C: 16.1
Falcon-7B	A_1: 30.1 A_2: 1.0 A_C: 0.0	A_1: 8.6 A_2: 2.1 A_C: 4.8	A_1: 39.7 A_2: 5.3 **A_C: 16.6**	A_1: 25.8 A_2: 4.2 A_C: 8.0	A_1: 26.8 A_2: 3.2 A_C: 11.2	A_1: 36.5 A_2: 3.2 A_C: 13.9	A_1: 30.1 A_2: 1.0 A_C: 0.0	**A_1: 52.6** A_2: 9.6 A_C: NA	A_1: 48.3 A_2: **19.3** A_C: 13.9

Table 3: AmbigQA-Cite Results. Accuracy scores are reported as percentages. The Document Split method involves providing each document individually to the models, and hence, citations are known by default. C.A=Conflict-aware.

Method / Model	Zero-Shot	C.A Basic Prompting			Few-shot C.A CoT		Zero-shot CoT	Document Split	Finetuning
		1-shot	3-shot	5-shot	1-shot	3-shot	1-shot		
Llama-7B	A_1: 84.6 A_2: 10.2 A_C: 0.0	A_1: 85.9 A_2: 64.0 A_C: 51.6	**A_1: 88.5** **A_2: 76.4** **A_C: 77.6**	A_1: 0.1 A_2: 0.0 A_C: 0.0	A_1: 81.7 A_2: 51.4 A_C: 14.4	A_1: 86.3 A_2: 68.7 A_C: 50.0	A_1: 81.5 A_2: 14.9 A_C: 0.0	A_1: 87.5 A_2: 49.0 A_C: NA	A_1: 79.3 A_2: 61.0 A_C: 58.5
Llama-13B	A_1: 82.2 A_2: 10.5 A_C: 0.0	A_1: 89.0 A_2: 74.5 A_C: 76.0	**A_1: 91.9** **A_2: 79.0** **A_C: 81.9**	A_1: 0.1 A_2: 0.0 A_C: 0.0	A_1: 86.8 A_2: 55.0 A_C: 23.6	A_1: 90.2 A_2: 74.3 A_C: 45.5	A_1: 80.7 A_2: 9.8 A_C: 0.0	A_1: 85.9 A_2: 40.0 A_C: NA	A_1: 81.6 A_2: 68.0 A_C: 68.1
Llama-70B	A_1: 88.3 A_2: 16.4 A_C: 0.0	A_1: 93.6 A_2: 85.1 A_C: 76.4	**A_1: 94.1** **A_2: 88.3** **A_C: 86.7**	A_1: 0.1 A_2: 0.0 A_C: 0.0	A_1: 92.3 A_2: 66.7 A_C: 26.7	A_1: 93.4 A_2: 83.6 A_C: 52.5	A_1: 75.8 A_2: 16.8 A_C: 0.0	A_1: 91.5 A_2: 45.8 A_C: NA	A_1: - A_2: - A_C: -
MPT-7B	A_1: 80.3 A_2: 2.7 A_C: 0.0	A_1: 78.2 A_2: 42.3 A_C: 43.1	A_1: 74.0 A_2: 49.3 **A_C: 54.1**	A_1: 0.1 A_2: 0.0 A_C: 0.0	A_1: 75.7 A_2: 35.7 A_C: 13.9	A_1: 70.9 A_2: 30.0 A_C: 10.0	A_1: 77.6 A_2: 4.8 A_C: 0.0	**A_1: 82.7** **A_2: 56.6** A_C: NA	A_1: 61.0 A_2: 21.0 A_C: 9.6
Falcon-7B	A_1: 63.2 A_2: 16.6 A_C: 0.0	A_1: 50.6 A_2: 35.6 A_C: 37.5	A_1: 70.3 **A_2: 45.8** **A_C: 53.0**	A_1: 0.0 A_2: 0.0 A_C: 0.0	A_1: 54.7 A_2: 30.0 A_C: 27.1	A_1: 69.9 A_2: 44.0 A_C: 42.0	A_1: 61.3 A_2: 10.8 A_C: 0.0	**A_1: 71.8** A_2: 38.3 A_C: NA	A_1: 71.6 A_2: 45.6 A_C: 46.3

Table 4: DisentQA-DupliCite Results. The Document Split method involves providing each document individually to the models, and hence, citations are known by default.

different string, yielding duplicated content. Comparing the results in Tables 4 and 6, we observe two significant trends: firstly, generating correct citations is much more challenging in the multi-hop setting, where multiple documents exist and are required to reach the answer. Secondly, producing all correct answers is also much harder, even with a limited number of correct ones. Moreover, the presence of distractors in the conflicting HotpotQA dataset further exacerbates this challenge, leading to an even more significant performance drop. These results underscore the importance of developing novel conflicting multi-hop QA datasets.

5.3 3-shot vs. 5-shot

While on the AmbigQA dataset we see a drop in performance across all models between the 3-shot and 5-shot few-shot method (see Table 3), the performance drop is far more significant in Table 4 on the DisentQA dataset. Analyzing this further, we find that with 5 examples the context becomes larger (especially on the DisentQA dataset) than the maximum context length the models can handle, which results in the test question truncation.

5.4 C.A Prompting vs. Zero-shot CoT

One possible reason for the zero-shot CoT's poor performance on A_C, with a score of 0% across all models and tested datasets, is that it lacks an explicit citation prompt. Unlike the C.A methods, which specifically ask models to cite their sources, the zero-shot method only generates a reasoning chain in the first step, without explicitly requesting citation. This highlights the necessity of a spe-

Method / Model	Zero-Shot	C.A Basic	Few-shot C.A CoT	Finetuning
Llama-7B	A_1: 69.6 A_2: 7.3 A_C: 0.0	A_1: 74.3 A_2: **56.0** A_C: **59.0**	A_1: 65.3 A_2: 47.6 A_C: 35.0	A_1: **74.6** A_2: 54.0 A_C: 40.0
Llama-13B	A_1: 71.6 A_2: 4.3 A_C: 0.0	A_1: 77.0 A_2: 58.0 A_C: 60.6	A_1: 72.0 A_2: 40.6 A_C: 10.8	A_1: **81.6** A_2: **66.6** A_C: **67.1**
MPT-7B	A_1: **65.3** A_2: 0.3 A_C: 0.0	A_1: 54.3 A_2: **26.6** A_C: 32.0	A_1: 56.0 A_2: 14.3 A_C: 4.3	A_1: 64.6 A_2: 22.6 A_C: **10.8**
Falcon-7B	A_1: 50.6 A_2: 7.6 A_C: 0.0	A_1: 54.3 A_2: 19.3 A_C: 30.3	A_1: 58.3 A_2: 22.0 A_C: 30.1	A_1: **69.3** A_2: **41.6** A_C: **40.1**

Table 5: DisentQA-ParaCite. C.A=Conflict-aware. We use 3 examples for both C.A Basic and CoT.

Method/ Model	Zero-Shot	C.A Basic	Few-shot C.A CoT	Finetuning
Llama-7B	A_1: 82.6 A_2: 27.6 A_3: 5.0 A_C: 0.0	A_1: 67.0 A_2: 36.1 A_3: 10.3 A_C: 8.9	A_1: 75.0 A_2: 25.0 A_3: 10.0 A_C: 11.6	A_1: **98.0** A_2: **90.0** A_3: **62.0** A_C: **67.3**
Llama-13B	A_1: 83.0 A_2: 21.3 A_3: 4.6 A_C: 0.0	A_1: 86.0 A_2: 68.9 A_3: 37.0 A_C: 36.3	A_1: 80.0 A_2: 55.0 A_3: 25.0 A_C: 13.3	A_1: **98.3** A_2: **93.3** A_3: **65.6** A_C: **76.3**
MPT-7B	A_1: 72.3 A_2: 16.0 A_3: 2.0 A_C: 0.0	A_1: 65.3 A_2: 24.0 A_3: 4.8 A_C: 0.03	A_1: 50.0 A_2: 15.0 A_3: 0.0 A_C: 0.0	A_1: **93.0** A_2: **84.3** A_3: **59.0** A_C: **64.5**
Falcon-7B	A_1: 63.0 A_2: 24.6 A_3: 6.3 A_C: 0.0	A_1: 48.1 A_2: 15.4 A_3: 2.6 A_C: 0.01	A_1: 0.0 A_2: 0.0 A_3: 0.0 A_C: 0.0	A_1: **85.3** A_2: **75.3** A_3: **39.3** A_C: **49.7**

Table 6: Conflicting HotpotQA-Cite (no distractors). C.A=Conflict-aware. We use 3 examples for both C.A Basic and CoT.

cific citation prompt. Furthermore, we observe a significant difference in A_2 scores between the two methods in both Tables 3 and 4, suggesting that models' self-generated reasoning chains are insufficient to handle conflicting facts.

5.5 Limited Efficacy of C.A. Prompts on HotpotQA

We find that both the C.A basic and C.A CoT perform worse than the zero-shot baseline and finetuning approach on the conflicting HotpotQA-Cite datasets. We hypothesize that this may be due to several reasons, such as the complexity of the multi-hop contexts, more cited documents in the multi-hop dataset, or that the in-context examples in the multi-hop setting were not as beneficial.

5.6 Finetuning vs. Prompting

Consistent with prior work (Brown et al., 2020; Liu et al., 2021; Wei et al., 2023; Dong et al., 2023), our results show that the C.A prompting method can achieve comparable or even better performance than finetuning on AmbigQA-Cite, DisentQA-ParaCite, and DisentQA-DupliCite. However, it struggles on Conflicting HotpotQA-Cite. Notably, finetuned models experience significant degradation when no conflicts exist. Overall, we conclude that the C.A basic method is the most effective approach, but both methods have room for improvement (see Section 7).

Method/ Model	Zero-Shot	C.A Basic	Few-shot C.A CoT	Finetuning
Llama-7B	A_1: **59.8** A_2: 16.8 A_3: 2.2 A_C: 0.0	A_1: 38.0 A_2: 9.3 A_3: 1.0 A_C: 0.1	A_1: 39.3 A_2: 12.6 A_3: 0.6 A_C: 0.2	A_1: 49.0 A_2: **17.3** A_3: **2.3** A_C: **2.0**
Llama-13B	A_1: **60.8** A_2: 15.8 A_3: 3.5 A_C: 0.0	A_1: 51.0 A_2: **20.0** A_3: **2.6** A_C: **1.3**	A_1: 42.3 A_2: 13.0 A_3: 2.3 A_C: 0.1	A_1: 46.6 A_2: 16.3 A_3: 2.0 A_C: 1.2
MPT-7B	A_1: **49.5** A_2: 13.0 A_3: **3.0** A_C: 0.0	A_1: 31.0 A_2: 8.6 A_3: 1.0 A_C: 0.0	A_1: 42.0 A_2: 12.3 A_3: 1.3 A_C: 0.0	A_1: 48.6 A_2: **15.3** A_3: 1.6 A_C: **0.4**
Falcon-7B	A_1: 29.5 A_2: 7.5 A_3: 2.5 A_C: 0.0	A_1: 25.3 A_2: 8.0 A_3: 1.3 A_C: 0.0	A_1: 4.3 A_2: 0.0 A_3: 0.0 A_C: 0.0	A_1: **37.4** A_2: **8.4** A_3: **8.6** A_C: **0.2**

Table 7: Conflicting HotpotQA-Cite (Distractors). C.A=Conflict-aware. We use 3 examples for both C.A Basic and CoT.

Method/ Model	Zero-Shot	C.A Basic	Finetuning
Llama-7B	**A_1: 84.8**	A_1: 84.1	A_1: 64.3
Llama-13B	**A_1: 83.8**	A_1: 80.3	A_1: 71.6
Llama-70B	**A_1: 89.0**	A_1: 88.1	A_1: -
MPT-7B	**A_1: 81.2**	A_1: 73.9	A_1: 51.3
Falcon-7B	A_1: 71.1	**A_1: 72.5**	A_1: 46.3

Table 8: DisentQA with no contextual conflicts. C.A=Conflict-aware. We use 3 examples for C.A Basic.

6 Real-world Usage

In our comprehensive analysis, we evaluate three main approaches to improve LLMs' ability to answer ambiguous questions with source citations: 1) prompt-based; 2) rule-based, and 3) fine-tuning-based. Notably, while the rule-based approach outperforms the other two in some occasions, as discussed in Section 3.4.2, it is incapable of answering questions that require complex reasoning across multiple documents, as it only sees one document at a time. To that end, we do not recommend using this approach when it is known that the data is of complex nature. But, to use it, users need to split the retrieved documents into chunks of one document at a time, which are sent to the model, followed by an aggregation of the answers. With regards to the other two approaches, the prompt-based approaches can be incorporated into most LLMs with a simple addition of a prompt, as shown in Appendix A. However, it is worth mentioning that the fine-tuning approach outperforms the prompting approach on multihop reasoning, but also results in a large performance decrease when no ambiguity exists, as discussed in Section 4.4. We also showed that LoRA-based fine-tuning is sufficient to improve LLMs' abilities in this task greatly over the baseline, highlighting the usability for real-users that do not have large computational resources.

7 Future Work

Having established the novel task of QA with source citation in ambiguous settings, and proposing strong baselines across our newly created datasets, several promising directions for future research emerge. Future work can build upon our contributions by: 1) Investigating the impact of finetuning on datasets with distractors, as our current experiments only focused on the no-distractor setting; 2) extending our finetuning data to include both conflicting and non-conflicting instances, which could lead to more robust models capable of handling varying levels of ambiguity; 3) exploring the effectiveness of finetuning on paraphrased data instead of duplications, which may provide additional insights into the model's ability to generalize across different linguistic formulations; 4) designing alternative prompts that emphasize the importance of citations or exploring other citation methods to further enhance the model's performance; 5) creating a more diverse range of datasets that capture different aspects of ambiguity, beyond the three flavors (paraphrasing, distractors, and duplications) we experiment with in this work; 6) investigating the potential benefits of employing different architectures or larger language models to tackle this challenging task; and 7) pretraining models on conflicting data and citations to potentially improve their ability to resolve knowledge conflicts and provide trustworthy answers.

8 Conclusion

We address a significant gap in QA research by introducing the novel task of QA with source citation in ambiguous settings. This task combines the complexities of ambiguous QA with the importance of providing source citations, enabling users to evaluate the factuality of each answer. To facilitate research in this area, we provide a comprehensive framework consisting of novel datasets, new evaluation metrics, and strong baselines using various approaches over 5 LLMs. Our work aims to inspire the development of more trustworthy and interpretable QA systems, bridging the gap between ambiguous answer resolution and citation generation. By exploring this new task, we hope to pave the way for more reliable and transparent QA systems that can accurately resolve knowledge conflicts and provide users with credible sources to support their answers.

Limitations

While we have evaluated 5 diverse LLMs and observed similar limitations, it is an open question whether other LLMs would face similar challenges in generating citations and handling multiple answers or ambiguous context. However, our findings suggest that many models may struggle with these tasks. Furthermore, our focus on unstructured texts from standard reading comprehension datasets raises the question of whether other knowledge formats, such as knowledge graph triples, would yield similar results.

Ethics Statement

Our motivation is to create systems that can effectively handle conflicting information and provide source citations, enabling users to verify the accuracy of the answers. We emphasize the importance of future research building upon this foundation to develop reliable systems that can be safely deployed in real-world applications.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Anthropic. *Model Card and Evaluations for Claude Models Anthropic*.

Yang Bai, Anthony Colas, and Daisy Zhe Wang. 2023. Mythqa: Query-based large-scale checkworthy claim detection through multi-answer open-domain question answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23. ACM.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Haowei Du, Quzhe Huang, Chen Zhang, and Dongyan Zhao. 2022. Knowledge-enhanced iterative instruction generation and reasoning for knowledge base question answering.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. 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.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Yifan Gao, Henghui Zhu, Patrick Ng, Cicero Nogueira dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Answering ambiguous questions through generative evidence fusion and round-trip prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3263–3276, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge.

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.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Dacheng Li, Rulin Shao*, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, , and Hao Zhang. 2023a. How long can open-source llms truly promise on context length?

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

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.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023a. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. Joint passage ranking for diverse multi-answer retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6997–7008, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020a. AmbigQA: Answering ambiguous open-domain questions. In *EMNLP*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020b. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Azade Mohammadi, Reza Ramezani, and Ahmad Baraani. 2022. A comprehensive survey on multi-hop machine reading comprehension approaches.

MPT. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

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

OpenAI. 2023b. Gpt-4 technical report.

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

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023. QACheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 264–273, Singapore. Association for Computational Linguistics.

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

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

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

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Sagi Shaier, Kevin Bennett, Lawrence Hunter, and Katharina von der Wense. 2024a. Comparing template-based and template-free language model probing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 766–776, St. Julian's, Malta. Association for Computational Linguistics.

Sagi Shaier, Lawrence Hunter, and Katharina Kann. 2022. Mind the knowledge gap: A survey of knowledge-enhanced dialogue systems.

Sagi Shaier, Lawrence Hunter, and Katharina Kann. 2023a. Who are all the stochastic parrots imitating? they should tell us! In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 113–120, Nusa Dua, Bali. Association for Computational Linguistics.

Sagi Shaier, Lawrence Hunter, and Katharina von der Wense. 2024b. It is not about what you say, it is about how you say it: A surprisingly simple approach for improving reading comprehension. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8292–8305, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sagi Shaier, Lawrence Hunter, and Katharina Wense. 2024c. Desiderata for the context use of question answering systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 777–792, St. Julian's, Malta. Association for Computational Linguistics.

Sagi Shaier, Lawrence E. Hunter, and Katharina von der Wense. 2023b. Who are all the stochastic parrots imitating? they should tell us!

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Zhihong Shao and Minlie Huang. 2022. Answering open-domain multi-answer questions via a recall-then-verify framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1838, Dublin, Ireland. Association for Computational Linguistics.

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.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation.

Weiwei Sun, Hengyi Cai, Hongshen Chen, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023. Answering ambiguous questions via iterative prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7669–7683, Toronto, Canada. 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. 2021. Knowledge enhanced pretrained language models: A compreshensive survey.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Michael J. Q. Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms.

A Prompts Used

A.1 Conflict-aware Basic Prompting

We use the following prompt design:

`You will get a question and some context texts. These texts may conflict with each other. If they do, answer according to the following examples: Example 1: Question:` q_{e1} `Context:` c_{e1} `Answer: According to Document D_1 the answer is A_1. According to Document D_2 the answer is A_2 [...] Question:` q_{ti} `Context:` c_{ti}`. Answer:`

A.2 Few-shot Conflict-aware CoT Prompting

`Example 1: Question:` q_{e1}`. Context:` c_{e1}`. Reasoning: Document D_1 mentions that the answer is A_1. But, Document D_2 mentions that the answer is A_2. Therefore, the answer is: According to Document D_1 the answer is A_1. According to Document D_2 the answer is A_2. Question:` q_{ti} `Context:` c_{ti}`. Answer:`

A.3 Zero-shot CoT Prompting

Step 1

`Question:` q_{ti} `Context:` c_{ti}`. Answer: Let's think step by step.`

This produces a generated response r_i.

Step 2

`Question:` q_{t1} `Context:` c_{t1}`.` r_i`. Therefore, the answer is`

B Finetuning Parameters

Our fine-tuning settings include: LoRA with alpha=15 and dropout=0.1; 1000 steps with early stopping (patience=3) guided by the validation set; batch size=4; gradient accumulation steps=1; AdamW optimizer with a linear learning rate scheduler and 10 warmup steps.