HALLUCINATION AUGMENTED RECITATIONS FOR LANGUAGE MODELS

Abdullatif Köksal[1,2*] **Renat Aksitov**[3] **Chung-Ching Chang**[3]
[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center of Machine Learning [3]Google Research
akoksal@cis.lmu.de {raksitov, ccchang}@google.com

ABSTRACT

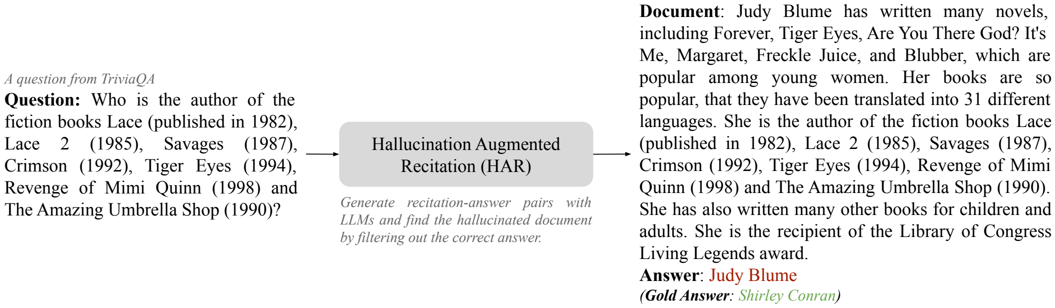
Attribution is a key concept in large language models (LLMs) as it enables control over information sources and enhances the factuality of LLMs. While existing approaches utilize open book question answering to improve attribution, factual datasets may reward language models to recall facts that they already know from their pretraining data. In contrast, counterfactual open book QA datasets would further improve attribution because the answer could only be grounded in the given text. We propose Hallucination Augmented Recitations (HAR) for creating counterfactual datasets by utilizing hallucination in LLMs for this purpose. For open book QA as a case study, we demonstrate that models finetuned with our counterfactual datasets improve text grounding, leading to better open book QA performance, with up to an 8.0% increase in F1 score. Our counterfactual dataset leads to significantly better performance than using human-annotated factual datasets, even with 4x smaller datasets and 4x smaller models. We observe that improvements are consistent across various model sizes and datasets, including multi-hop, biomedical, and adversarial QA datasets.

1 INTRODUCTION

Text grounding or attribution[1] is a key aspect in large language models (LLMs). Since most LLMs are trained once without any update, attributability gives LLMs an adaptation ability to dynamic changes in the real world, such as temporal questions (Vu et al., 2023). Additionally, attribution improves the factuality of language models and could help to control the source of information more granularly (Gao et al., 2023). Recent works in LLMs focus on adapting retrieval-augmented approaches, such as search-engine-assisted systems like Bard and GPT-4, to utilize attributable LLMs (Lewis et al., 2020; Chen et al., 2017).

To improve the text grounding abilities of language models, most works focus on including open book question answer (QA) tasks which include attributable documents for given questions. Therefore, the largest tasks in the training datasets of recent instruction-tuned models such as T*k*-Instruct (Wang et al., 2022b) or InstructGPT (Ouyang et al., 2022) are based on open book QA. However, there is an underlying multi-objective trade-off in open book QA. Since pretrained language models already know facts through their pretraining data, finetuning them with factual open book QA datasets could reward the model to recall the facts from their memory without attribution instead of attributing to the document. Therefore, attribution and recall are competing factors in language models because facts in open book QA datasets may already present in the memory of language models. In contrast, counterfactual data mitigate such spurious correlations and further improve the text grounding abilities of language models (Kaushik et al., 2020). This is because counterfactual data introduces a conflict with the memory, preventing straightforward recall. However, recent work focusing on counterfactual open book QA via entity substitution or retrieval-based generation demonstrates only limited and inconsistent improvements in text grounding (Longpre et al., 2021; Paranjape et al., 2022).

[*] Work performed during an internship at Google.
[1] We use these terms interchangeably throughout the paper.


A question from TriviaQA
Hallucination Augmented Recitation (HAR)
Generate recitation-answer pairs with LLMs and find the hallucinated document by filtering out the correct answer.

Question: Who is the author of the fiction books Lace (published in 1982), Lace 2 (1985), Savages (1987), Crimson (1992), Tiger Eyes (1994), Revenge of Mimi Quinn (1998) and The Amazing Umbrella Shop (1990)?

Document: Judy Blume has written many novels, including Forever, Tiger Eyes, Are You There God? It's Me, Margaret, Freckle Juice, and Blubber, which are popular among young women. Her books are so popular, that they have been translated into 31 different languages. She is the author of the fiction books Lace (published in 1982), Lace 2 (1985), Savages (1987), Crimson (1992), Tiger Eyes (1994), Revenge of Mimi Quinn (1998) and The Amazing Umbrella Shop (1990). She has also written many other books for children and adults. She is the recipient of the Library of Congress Living Legends award.
Answer: Judy Blume
(Gold Answer**: Shirley Conran**)

Figure 1: A counterfactual example from CF-TriviaQA, generated via the HAR pipeline. For a given question from TriviaQA, we utilize LLM hallucination to generate high-quality, attributable, and counterfactual open book QA examples. As illustrated in this example, our HAR pipeline outputs a hallucinated[3] document supporting the counterfactual answer of Judy Blume, while the gold answer is Shirley Conran as given in TriviaQA.

We propose **Hallucination Augmented Recitations (HAR)** which utilizes LLM hallucination[2] to create a counterfactual open book QA dataset. HAR builds on the recitation augmentation (Sun et al., 2023) by prompting LLMs to introduce reasoning through recitation and produce an attributable document for a given question. This produces high-quality and consistent counterfactuals, in contrast to heuristics such as entity replacement (Longpre et al., 2021). We then apply additional filters to identify high-quality hallucinations from LLMs and create a counterfactual dataset. In Figure 1, we illustrate an example from our dataset, CF-TriviaQA. For a given question about an author of several books, the LLM hallucinates and generates a counterfactual document that supports an incorrect answer (Judy Blume). Therefore, a model finetuned on such counterfactual open book QA dataset can be rewarded only by attributing to the document, since the answer cannot be recollected from the model's memory.

Our contributions are as follows:

1. We utilize hallucination and propose the **Hallucination Augmented Recitations (HAR)** pipeline to create a high-quality attributable counterfactual open book QA, named **CF-TriviaQA** with 19K examples.
2. We show that T5 models finetuned with CF-TriviaQA significantly outperform those finetuned with human-annotated factual open book QA datasets, even **with 4x smaller datasets and 4x smaller model sizes**.
3. We observe that our findings are consistent across various model sizes, ranging **from 60M to 11B**, and on various datasets, including **multi-hop**, **biomedical**, or **adversarial** questions.

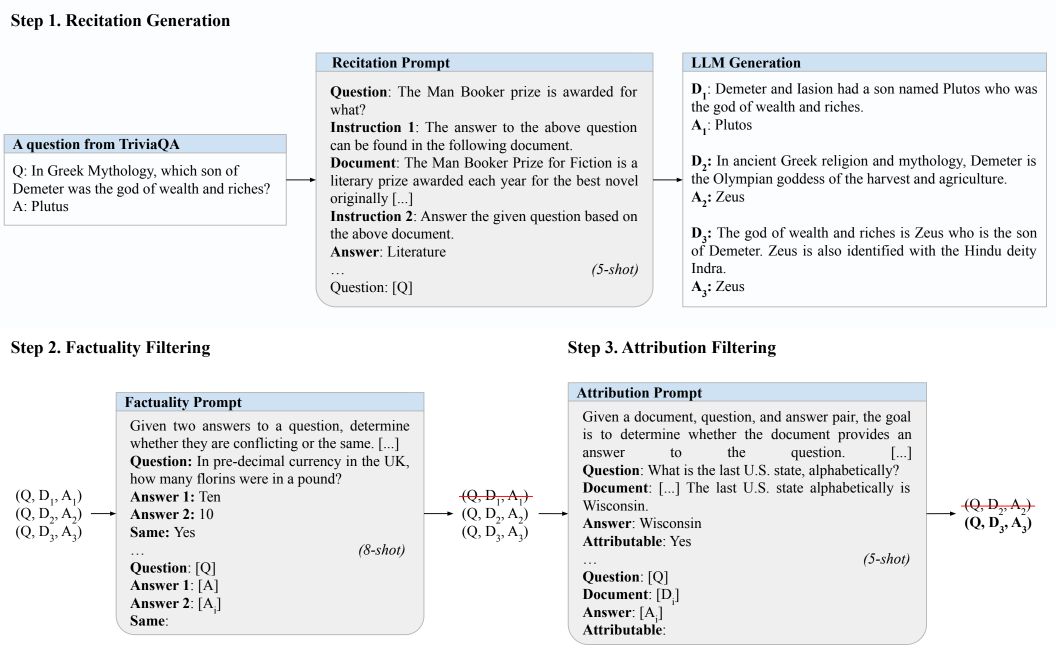
2 Hallucination Augmented Recitations

We aim to create a counterfactual question answering dataset to further improve attribution in language models. To this end, we utilize hallucination for counterfactuals and propose Hallucination Augmented Recitations (HAR). HAR has three steps as illustrated in Figure 2.

1. **Recitation generation**: We use the recitation-augmented language model approach (Sun et al., 2023) to generate multiple document and answer pairs for a given question.
2. **Factuality Filtering**: We filter out factual answers to focus on counterfactuality.
3. **Attribution Filtering**: We apply the attribution filter to remove question, document, and answer pairs where the answer is not grounded in the document.

[2] We define hallucination as the generation of counterfactual text conflicting with real-world knowledge.

[3] LLM's hallucination in this example is likely due to the ambiguous book name "Tiger Eyes", as both Judy Blume and Shirley Conran have authored books with that title. However, Judy Blume's Tiger Eyes was published in 1981, and she did not write any books with the other names listed in the question.



Q: In Greek Mythology, which son of Demeter was the god of wealth and riches?
A: Plutus

Question: The Man Booker prize is awarded for what?
Instruction 1: The answer to the above question can be found in the following document.
Document: The Man Booker Prize for Fiction is a literary prize awarded each year for the best novel originally [...]
Instruction 2: Answer the given question based on the above document.
Answer: Literature
… *(5-shot)*
Question: [Q]

D_1: Demeter and Iasion had a son named Plutos who was the god of wealth and riches.
A_1: Plutos

D_2: In ancient Greek religion and mythology, Demeter is the Olympian goddess of the harvest and agriculture.
A_2: Zeus

D_3: The god of wealth and riches is Zeus who is the son of Demeter. Zeus is also identified with the Hindu deity Indra.
A_3: Zeus

(Q, D_1, A_1)
(Q, D_2, A_2)
(Q, D_3, A_3)

Given two answers to a question, determine whether they are conflicting or the same. [...]
Question: In pre-decimal currency in the UK, how many florins were in a pound?
Answer 1: Ten
Answer 2: 10
Same: Yes
… *(8-shot)*
Question: [Q]
Answer 1: [A]
Answer 2: [A_i]
Same:

~~(Q, D_1, A_1)~~
(Q, D_2, A_2)
(Q, D_3, A_3)

Given a document, question, and answer pair, the goal is to determine whether the document provides an answer to the question. [...]
Question: What is the last U.S. state, alphabetically?
Document: [...] The last U.S. state alphabetically is Wisconsin.
Answer: Wisconsin
Attributable: Yes
… *(5-shot)*
Question: [Q]
Document: [D_i]
Answer: [A_i]
Attributable:

~~(Q, D_2, A_2)~~
(Q, D_3, A_3)

Figure 2: Three steps of HAR. HAR first generates document, and answer pairs for a given question. Then, it filters factual generations (e.g., the answer with Plutos while the gold answer is Plutus). Finally, it filters generated examples without text grounding (e.g., the second document-answer pair, where there is no mention of Zeus in the document). See full prompts in §B.

2.1 Recitation Generation

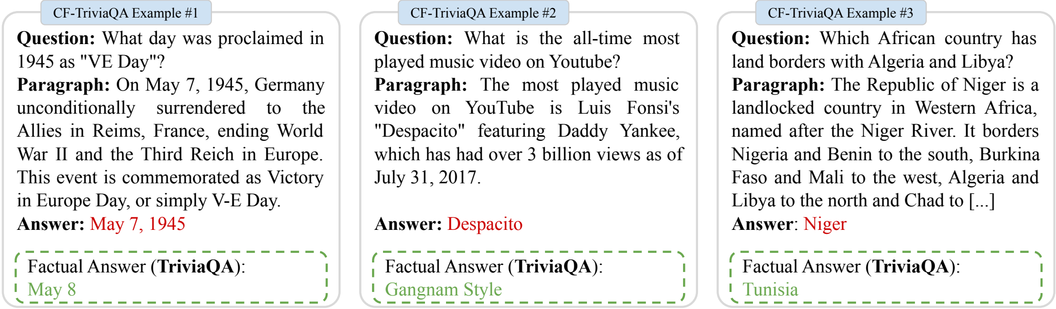
Recitation generation is the first step of HAR, as described in Figure 2. We follow Sun et al. (2023) and apply the recitation-augmented language model approach. Simply, we generate multiple document and answer pairs for a given question by using a 5-shot prompt via LLMs to create open book QA examples. We design the recitation prompt as shown in Figure 5 in the Appendix which conditions LLMs to generate document, and answer pairs for a given question. We manually pick five high-quality examples from TriviaQA (Joshi et al., 2017) as our few-shot examples.

During generation, we use questions only from TriviaQA (Joshi et al., 2017), and utilize PaLM 2-L (Anil et al., 2023) as our LLM in recitation generation. For each question, we perform 24 iterations with temperature sampling ($T = 0.7$). Then, we parse generated documents and answers and eliminate examples that do not follow our prompt (e.g., no new line between document and instruction 2).

2.2 Factuality Filtering

Since we only focus on counterfactual examples to improve attribution, we filter out examples that have factual answers. First, we simply remove generated examples whose answers have the same surface form as the gold answer in TriviaQA. However, our manual evaluation of the surface form filtering shows that there are many generated pairs where the generated answers are factual but their surface forms are different from the gold answers. It could be caused by synonyms, transliterations, accented characters, or reformatted answers (e.g., 1930 vs. 30s, Eugène Delacroix vs. Eugene Delacroix, or Plutos vs. Plutus as in Figure 2); therefore heuristics that apply filtering on surface forms are not enough to filter factual answers. In the second part of HAR, we propose a factuality filtering method via LLMs to remove such factual answers and keep counterfactual answers only.

We propose a PaLM 2-L based filtering method with 8-shot examples. In the prompt, we only provide question and answer pairs without a document because we are only interested in finding whether the generated answer leads to the same answer as the gold answer for the given question. We



CF-TriviaQA Example #1

Question: What day was proclaimed in 1945 as "VE Day"?
Paragraph: On May 7, 1945, Germany unconditionally surrendered to the Allies in Reims, France, ending World War II and the Third Reich in Europe. This event is commemorated as Victory in Europe Day, or simply V-E Day.
Answer: May 7, 1945

Factual Answer (**TriviaQA**):
May 8

CF-TriviaQA Example #2

Question: What is the all-time most played music video on Youtube?
Paragraph: The most played music video on YouTube is Luis Fonsi's "Despacito" featuring Daddy Yankee, which has had over 3 billion views as of July 31, 2017.

Answer: Despacito

Factual Answer (**TriviaQA**):
Gangnam Style

CF-TriviaQA Example #3

Question: Which African country has land borders with Algeria and Libya?
Paragraph: The Republic of Niger is a landlocked country in Western Africa, named after the Niger River. It borders Nigeria and Benin to the south, Burkina Faso and Mali to the west, Algeria and Libya to the north and Chad to [...]
Answer: Niger

Factual Answer (**TriviaQA**):
Tunisia

Figure 3: Three examples from CF-TriviaQA. They are counterfactual and conflict with the answers in TriviaQA, yet they have different attributes. The first example is simply counterfactual with an incorrect date, while the second example includes a temporal question in which the answer could change over time. The last example is an ambiguous question since there are two factual answers but the paragraph provides only one response.

select four examples whose generated answers are the same as the gold answers but have different surface forms and four examples with different answers. We design the prompt as illustrated in Figure 6 in the Appendix. We compare the probabilities of ‘Yes’ and ‘No’ tokens for the next token and decide factuality based on the normalized probability of the ‘Yes’ token.

2.3 Attribution Filtering

After the recitation generation and factuality filtering steps, we have a set of question, document, and answer pairs with only counterfactual answers. During our manual analysis of these examples, we encounter examples where the generated answer is not grounded in the generated document. In order to eliminate such cases, we again use filtering via LLMs (i.e., PaLM 2-L) with 5-shot examples. We design a prompt with generated counterfactual examples with and without grounding, as shown in Figure 7 in the Appendix.

We calculate the probability of ‘Yes’ and ‘No’ tokens and normalize it. If the normalized probability of the ‘Yes’ token is lower than 0.5, we remove those samples from our dataset. Furthermore, we may have different document, and answer pairs for a given question from the HAR pipeline since we generate 24 examples for the recitation generation part. To select a unique sample for each question, we select the document answer pair with the highest normalized attribution score.

3 Counterfactual TriviaQA: CF-TriviaQA

We propose CF-TriviaQA, a counterfactual dataset generated from the TriviaQA dataset (Joshi et al., 2017) using Hallucination Augmented Recitations (HAR). To construct CF-TriviaQA, we first apply recitation generation to generate 24 document, answer pairs for each question in TriviaQA, resulting in an average of approximately 3 unique answers per question. More than 30% of the generated examples have the gold answer as a generated answer, which means they are not counterfactual. Furthermore, we observe that generated answers with different surface forms than the gold answer could be still factual (e.g., synonyms, hyponymy, translation), or the generated answer may not be grounded in the generated text.

Next, we apply the factuality filter to remove factual answers, which eliminates more than 45% of the remaining examples. We then apply an attribution filter to remove generations without text grounding, which also removes more than 50% of the remaining data after the factuality filtering. We also select only one context for each answer with the highest attribution score according to our filter. Finally, we obtain our counterfactual dataset, **CF-TriviaQA**, with 19,327 examples.

CF-TriviaQA contains different types of counterfactual examples, as illustrated in Figure 3. The first example shows a simple counterfactual example with a different date than the factual date in the gold data. We mostly see examples in this category in CF-TriviaQA, simply counterfactual. The second example illustrates the temporal aspect of counterfactuality. The gold answer was factual when TriviaQA was first published, but due to the temporal aspect of the question, the original gold answer is no longer factual. Our generated example is also counterfactual since the gold answer has also changed over time, but CF-TriviaQA and TriviaQA examples conflict with each other. The final example showcases ambiguous questions. For this question, we could consider both Niger and Tunisia as potential answers, as they have land borders with Algeria and Libya. However, only the answer 'Niger' is attributable because the generated document does not mention Tunisia. The second and third examples illustrate another aspect of our HAR pipeline: HAR can produce conflicting open book QA examples for a given dataset.

3.1 Evaluation

We evaluate CF-TriviaQA in two aspects: attribution and counterfactuality. Following prior work (Rashkin et al., 2023; Honovich et al., 2022), we utilize natural language inference (NLI) tasks for evaluation. We use a T5-11B model finetuned with a mixture of NLI, fact verification, and paraphrase detection datasets, MNLI, SNLI, FEVER, PAWS, SciTail, and VitaminC, as proposed in Gao et al. (2023). We follow the NLI formulation for open book QA[4] in Chang et al. (2023); Aksitov et al. (2023) and measure attribution and counterfactuality scores as follows:

Attribution: We measure the entailment score when the premise is the generated document and question, and the hypothesis is the question and the generated answer. A high attribution score means that the generated answer is grounded in the generated document.

Counterfactuality: We measure the *contradiction* score when the premise is the generated document and question, and the hypothesis is the question and the *gold answer* in TriviaQA. Since we want CF-TriviaQA to be counterfactual, we would like there to be no entailment between the generated counterfactual document and the original factual gold answer.

We present the results in Table 1 before and after the filtering steps. Without any filtering, the dataset includes many factual examples without text grounding. However, factuality filtering improves counterfactuality by 0.19, and attribution filtering improves overall attribution by 0.12 points. These qualitative examples and improvements in NLI-based attribution and counterfactuality scores show the importance of the filtering mechanism.

	Attribution	Counterfactuality
CF-TriviaQA	**0.77**	**0.87**
- Attribution Filtering	0.65	0.84
- Factuality Filtering	0.68	0.65

Table 1: Attribution and counterfactuality evaluation of CF-TriviaQA via NLI. We show that each step of filtering improves the attribution and counterfactuality of our dataset.

In addition to attribution and counterfactuality scores presented in Table 1, we present qualitative examples labeled as factual by the counterfactual filter in Table 6 in the Appendix. These examples illustrate that the surface form heuristics to detect counterfactuals would not be sufficient. We observe a diverse set of factual answer generations that do not have the same surface form as the gold answer, such as hypernyms, ancient names/synonyms, or round numbers. We also illustrate generated examples without grounding according to the attribution filter in Table 7 in the Appendix. Although they are all counterfactual examples, they do not have a proper grounding in the text, or there is conflicting information between the question, generated document, and/or generated answer.

4 Open book QA Experiments

After describing our experimental setup, we ask 4 important research questions to analyze and measure the effect of counterfactual examples on text grounding by focusing on open book QA.

We finetune T5 models (Raffel et al., 2020) on either TriviaQA (Joshi et al., 2017), referred to as factual models, or on CF-TriviaQA, referred to as counterfactual models, or on the combination of TriviaQA and CF-TriviaQA, which is called the combined model. We follow the MRQA format (Fisch et al., 2019) in which TriviaQA includes 73K training samples while CF-TriviaQA includes 19K training samples. During evaluation, their performance is compared across *out-of-domain datasets*, following the experimental setup in Paranjape et al. (2022). The evaluation includes a comparison

[4] premise: {document}\n\n{question} hypothesis: {question}\n{answer}

Training Dataset	TriviaQA	SQuAD	NQ	HotpotQA	BioASQ	AQA	AmbigQA	OOD Avg.
TriviaQA	**85.2**	79.6	66.5	69.4	63.4	42.1	**53.2**	62.4
CF-TriviaQA	81.7	**81.7**	**71.2**	**73.8**	**69.5**	**44.9**	**53.2**	**65.7**

Table 2: Token-level F1 scores of T5-3B models finetuned with TriviaQA vs. CF-TriviaQA. T5-3B model finetuned with CF-TriviaQA significantly outperforms T5-3B with TriviaQA by 3.3 points.

across various open book QA datasets, including SQuAD (11K samples) (Rajpurkar et al., 2016), Natural Questions (4K samples) (NQ; Kwiatkowski et al., 2019), HotpotQA (6K samples) (Yang et al., 2018), BioASQ (2K samples) (Tsatsaronis et al., 2015), AmbigQA (1K samples) (Min et al., 2020), and adversarial QA (1K samples) (AQA; Bartolo et al., 2020) with the versions of MRQA 2019 shared task (Fisch et al., 2019). We mainly present the results with token-level F1 scores, measuring the partial match of predicted answers with gold answers, following prior work. Furthermore, we include exact match scores in §C in the Appendix, which exhibit similar trends to token-level F1 scores.

Q1. Does the (hallucination augmented) counterfactual dataset improve text grounding?

We finetune T5-3B models with human-annotated factual TriviaQA and counterfactual CF-TriviaQA with Hallucination Augmented Generation (HAR). We compare their performance in various out-of-domain datasets to see the effect of counterfactuals on text grounding.

Training Dataset	TriviaQA	OOD Avg.
TriviaQA	85.2	62.4
CF-TriviaQA	81.7	**65.7**
TriviaQA +CF-TriviaQA	**85.3**	65.3

Table 3: Token-level F1 scores of T5-3B models finetuned with TriviaQA, CF-TriviaQA, and their combination. Combining our CF-TriviaQA dataset with TriviaQA achieves good out-of-domain performance while having a similar performance in in-domain as the model finetuned with TriviaQA.

The results in Table 2 show that the counterfactual model achieves a **3.3** higher token-level F1 score with 4x smaller data than the factual model. It consistently outperforms the factual model on all out-of-domain datasets including multihop, biomedical, and adversarial QA datasets. We see a drop in in-domain performances as expected since the generated dataset conflicts with answers in TriviaQA and our focus is on generalization of text grounding abilities. However, the combined model, which is finetuned on both counterfactual CF-TriviaQA and factual TriviaQA, achieves both good in-domain performance and out-of-domain performance, as shown in Table 3.

Q2. Does the improvement in text grounding via counterfactuals vary with model size?

We observe a 5.3% relative improvement between T5-3B models finetuned with CF-TriviaQA and TriviaQA in Q1. To see how performance improvements vary across different model sizes, we finetune all T5 models, small (60M), base (220M), large (770M), 3B, and 11B.

We show the average F1 score in out-of-domain datasets in Figure 4. We see consistent improvements across all model sizes for both the counterfactual model with CF-TriviaQA and the combined model with CF-TriviaQA and TriviaQA over the factual model with TriviaQA. The relative improvement is always between 4.5% and 8.0% across all model sizes, suggesting that counterfactuals help to improve language models' text grounding capabilities regardless of the model size.

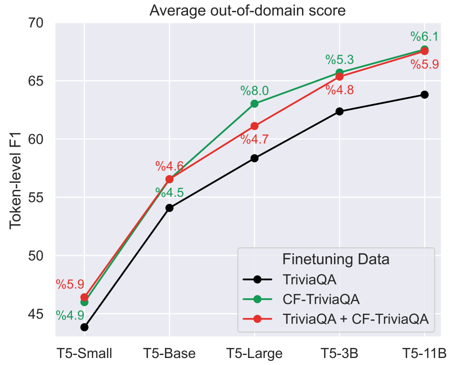


Figure 4: Out-of-domain performance of factual, counterfactual, and combined models with all sizes of T5 models. Models including counterfactual examples consistently outperform factual models across all sizes.

Furthermore, we can observe that the counterfactual models could achieve much better performance than the larger factual models. In Figure 4, we observe that T5-3B with CF-TriviaQA outperforms 4x larger model T5-11B with a 4x larger training dataset, with TriviaQA. This verifies the impact of both our HAR approach to generate counterfactual examples and counterfactuals on text grounding.

Q3. What would be the impact on text grounding performance if a factual open book QA dataset, generated through our modified HAR pipeline, was used for finetuning instead of using a counterfactual dataset?

Our HAR pipeline includes two novel contributions: LLM-generated open book QA dataset, and, more importantly, utilizing hallucination to generate counterfactual examples. To analyze the impact of using an LLM-generated dataset, we generate a factual alternative and compare their performance on out-of-domain open book QA datasets.

We modify the hallucination augmented retrieval pipeline to generate factual open book QA examples which is called F-TriviaQA. After recitation generation, we select factual recitations by comparing the generated answer to the gold answer instead of performing factuality filtering in HAR (i.e., step 2). We still perform attribution filtering (i.e., step 3) with F-TriviaQA to get high-quality grounded examples as well. To minimize the impact of other factors, we only include the same questions in CF-TriviaQA to F-TriviaQA. However, there are only around 10K questions common both in CF-TriviaQA and F-TriviaQA due to filtering, therefore we select around 9,000 generated factual questions randomly to keep them the same size.

Training Dataset	TriviaQA	OOD Avg.
TriviaQA	**85.2**	62.4
F-TriviaQA	83.3	63.6
CF-TriviaQA	81.7	**65.7**

Table 4: Token-level F1 scores of T5-3B model finetuned with TriviaQA (human-annotated factual), F-TriviaQA (LLM-generated factual), CF-TriviaQA (LLM-generated counterfactual). While F-TriviaQA outperforms TriviaQA, showing the strength of LLM generation and our HAR pipeline, CF-TriviaQA outperforms both factual models, showing the importance of counterfactuals on text grounding. See Table 10 in the Appendix for scores in each out-of-domain dataset, separately.

We compare the performance of T5-3B finetuned with TriviaQA (human-annotated factual), F-TriviaQA (LLM-generated factual), and CF-TriviaQA (LLM-generated counterfactual) on out-of-domain QA datasets. We present the results in Table 4. We observe that models finetuned with LLM-generated datasets (factual or counterfactual) based on HAR outperform the TriviaQA-based model even with a 4x smaller dataset. This finding is aligned with previous work showing improvements in generating synthetic data using LLMs (Wang et al., 2023; Köksal et al., 2023).

Furthermore, we see much higher improvement from the counterfactual dataset than from the LLM-generated factual dataset when both are compared with TriviaQA. The counterfactual model has a 3.3 higher token-level F1 score, while the model with F-TriviaQA has only a 1.2 higher token-level F1 score. This supports our hypothesis that counterfactuals improve text grounding.

Q4. What is the impact of LLM size in Hallucination Augmented Recitations (HAR)? Do smaller models generate the same-quality datasets that lead to similar improvements in text grounding?

We perform an additional study on the size of LLMs in HAR. We utilize LLM hallucination in the recitation generation of HAR and then filter to get high-quality examples. We replace PaLM 2-Large with a smaller variant, PaLM 2-Small for the recitation generation step while keeping PaLM 2-Large for filtering steps. We generate a new counterfactual dataset with the smaller PaLM model, called CF-TriviaQA$_{\text{PaLM 2-S}}$, which has the same number of examples as CF-TriviaQA$_{\text{PaLM 2-L}}$.

We observe that the hallucination rate in the smaller model is much higher than in the larger model. The initial dataset generated with PaLM 2-S has 31K high-quality counterfactual open book QA examples, which is 50% more than the data generated by PaLM 2-L. This is consistent with the previous work showing that smaller models tend to hallucinate more than their larger counterparts (Elaraby et al., 2023; Rawte et al., 2023). For a fair comparison, we again randomly sample the same number of examples (by including the common questions first) as CF-TriviaQA$_{\text{PaLM 2-L}}$.

We compare their performance on open book QA datasets in Table 5. The counterfactual model with CF-TriviaQA$_{\text{PaLM 2-S}}$ achieves even slightly better performance than CF-TriviaQA$_{\text{PaLM 2-L}}$. This

Training Dataset	**TriviaQA**	**OOD Avg.**
TriviaQA	**85.2**	62.4
CF-TriviaQA$_{\text{PaLM 2-S}}$	80.8	**66.0**
CF-TriviaQA$_{\text{PaLM 2-L}}$	81.7	65.7

Table 5: Comparison of LLM sizes in the HAR pipeline shows that smaller alternatives of PaLM 2 can achieve similar performance in out-of-domain scores. Therefore, smaller models can be used to utilize hallucination in HAR for better text grounding. See Table 12 in the Appendix for scores in each out-of-domain dataset, separately.

shows that the hallucination generated by smaller language models can be also used for counterfactual data generation via HAR after applying factuality and attribution filtering steps with larger models.

5 Related Work

Counterfactual Datasets: Counterfactuals in NLP usually refer to perturbations that make the given text true under different circumstances, while remaining consistent with the possible worlds where the prerequisites hold. Therefore, counterfactuals play a vital role in both the evaluation of language models (Qin et al., 2019; Wu et al., 2023) and their out-of-domain generalization (Bowman & Dahl, 2021) of language models. Prior works on counterfactual generation utilize expensive human annotation (Kaushik et al., 2020), while more recent works focus on automatic generation. Some of these works employ basic heuristics such as negating verbs or swapping noun phrases (Dua et al., 2021), and replacing gendered words in questions (Webster et al., 2020). Wang et al. (2022a) introduce a pipeline to generate synthetic negative summaries to improve faithfulness in abstractive text summarization. Paranjape et al. (2022) propose a retrieval-based generation system to create counterfactual datasets.

There are recent works focusing on perturbing contexts in open book QA with methods such as named entity replacement, thereby changing the answer to create counterfactual examples (Longpre et al., 2021; Ye et al., 2021). However, these methods have difficulty understanding complex structures and may create counterfactual examples that are not consistent (e.g., changing a date of birth without changing the date of death or without changing the occurrence of age in the document). This is reflected in their results, as these approaches have shown only weak and inconsistent improvements in open book QA (Paranjape et al., 2022; Longpre et al., 2021). In contrast, we propose LLM-based Hallucination Augmented Recitations (HAR) for counterfactual generation. HAR produces high-quality and consistent counterfactual examples, as seen by qualitative examples and out-of-domain performance improvement. To the best of our knowledge, we are the first ones to utilize LLM hallucination to create counterfactual datasets.

Synthetic Data Generation: Synthetic question answering dataset generation without counterfactuals has shown limited improvement in out-of-domain generalization (Bartolo et al., 2021; Lewis et al., 2021). However, recent advancements in large language models (LLMs) have led to growing interest in synthetic data generation with LLMs, such as in more generalized instruction tuning datasets from scratch (Wang et al., 2023) or by restructuring existing corpora (Köksal et al., 2023). Synthetic data generation with LLMs has also been applied to existing datasets for specific tasks to improve model quality, such as natural language inference (Liu et al., 2022) and sentiment analysis (Meng et al., 2023).

6 Conclusion

In this paper, we propose Hallucination Augmented Recitations (HAR) to create a counterfactual open book QA dataset, CF-TriviaQA. Since factual open book QA tasks have multi-objective tradeoffs (i.e., recalling the answer from the memory of language models vs. grounding in the given context), we hypothesize that high-quality counterfactual datasets would further improve attribution. Our results show that models finetuned with CF-TriviaQA significantly outperform models finetuned with factual TriviaQA, even with a 4x smaller training dataset and a 4x smaller model size. This improvement is consistent across various out-of-domain open book QA tasks, including multi-hop, biomedical, and adversarial questions. For future work, examples in CF-TriviaQA could help to analyze LLM hallucination as our HAR pipeline enables the generation of high-quality and complex hallucinations. Additionally, the generated counterfactual dataset and the HAR pipeline could be further used to evaluate the text grounding abilities of LLMs (Wu et al., 2023) or improve the robustness of natural language inference models, which could later be applied to improve attribution scoring (Rashkin et al., 2023).

7 Limitations

This paper presents a methodology to create a high-quality and attributable counterfactual open book QA dataset. While this counterfactual dataset has the potential to improve the text grounding and generalization abilities of language models, finetuning on counterfactual examples may have a negative impact on model factuality. Since our main focus is on improving attribution, we do not perform any analysis on potential impacts to factuality. Therefore, we recommend careful consideration before finetuning models on any counterfactual datasets created using our hallucination augmented recitations (HAR) pipeline. Furthermore, our filtering steps have demonstrated significant improvements in attribution and counterfactuality (see Table 1). It is crucial to acknowledge that LLMs can still make mistakes. As indicated in Tables 6 and 7, some generated examples may still have shortcomings despite our filtering methodology. Therefore, it is important to consider the limitations of LLMs when utilizing them for filtering since they may fail to identify and exclude subpar examples.

References

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models, 2023.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL https://aclanthology.org/2020.tacl-1.43.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8830–8848, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.696. URL https://aclanthology.org/2021.emnlp-main.696.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. Kl-divergence guided temperature sampling, 2023.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://aclanthology.org/P17-1171.

Dheeru Dua, Pradeep Dasigi, Sameer Singh, and Matt Gardner. Learning with instance bundles for reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7347–7357, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.584. URL https://aclanthology.org/2021.emnlp-main.584.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. Halo: Estimation and reduction of hallucinations in open-source weak large language models, 2023.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801. URL https://aclanthology.org/D19-5801.

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. 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)*, pp. 16477–16508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pp. 161–175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.19. URL https://aclanthology.org/2022.dialdoc-1.19.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 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)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

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. Natural questions: A benchmark for question answering research. *Transactions of the*

Association for Computational Linguistics, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. Longform: Optimizing instruction tuning for long text generation with corpus extraction, 2023.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115, 10 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00415. URL https://doi.org/10.1162/tacl_a_00415.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6826–6847, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.508. URL https://aclanthology.org/2022.findings-emnlp.508.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1670–1686, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.117. URL https://aclanthology.org/2022.acl-long.117.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference* *on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5043–5053, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1509. URL `https://aclanthology.org/D19-1509`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, pp. 1–64, 08 2023. ISSN 0891-2017. doi: 10.1162/coli_a_00486. URL `https://doi.org/10.1162/coli_a_00486`.

Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=-cqvvvb-NkI`.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, Apr 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0564-6. URL `https://doi.org/10.1186/s12859-015-0564-6`.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. Improving faithfulness by augmenting negative summaries from fake documents. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11913–11921, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.816. URL `https://aclanthology.org/2022.emnlp-main.816`.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL `https://aclanthology.org/2022.emnlp-main.340`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *CoRR*, abs/2010.06032, 2020. URL https://arxiv.org/abs/2010.06032.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Xi Ye, Rohan Nair, and Greg Durrett. Connecting attributions and QA model behavior on realistic counterfactuals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5496–5512, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.447. URL https://aclanthology.org/2021.emnlp-main.447.

Question: The element of the Periodic Table LUTETIUM has its name derived from the Latin name for the city in which is was discovered. Which city?	**Generated Document:** Lutetium is a chemical element with symbol Lu and atomic number 71. It is a silvery white metal which resists corrosion in dry, but not moist, air. It is the last element in the lanthanide series, and traditionally counted among the rare earths. Lutetium is usually considered to be the final element in the lanthanide series. However, lutetium has a 4f145d16s2 configuration, which makes it technically the first element of the 5d transition series. Lutetium was named after Lutetia, the Roman name of Paris, where its discovery was announced by Georges Urbain.	**Generated Answer** Lutetia, Paris	**Gold Answer** Paris
Notes: The generated answer includes the ancient name of Paris, Lutetia, likely to provide more insight into the question about Lutetium.			
Question: What was the population of the world at 31 Oct 2011 according to the United Nations?	**Generated Document:** The world population has been increasing rapidly. Between 1900 and 2000 it increased almost four times. Now it stands at about 6.5 billion. The current projection is that by the year 2025 it will be about 8 billion. United Nations estimates show that the world population at 31 Oct 2011 was 7,001,818,322.	**Generated Answer** 7,001,818,322	**Gold Answer** 7 billion
Notes: The generated answer provides more detailed information, but both answers are considered to be the same.			
Question: The introduction of what vegetable coincided with the coronation of the British monarch in 1902?	**Generated Document:** The year 1902 saw the coronation of King Edward VII and the introduction of the potato crisp to the British public. The crisp is not only the nation's favourite snack, it is also one of the few British inventions to make it big around the world. In 2004, Britain celebrated the centenary of the crisp, which was the brainchild of a frustrated chef.	**Generated Answer** Potato	**Gold Answer** King Edward potato
Notes: The generated answer is less specific than the gold answer, but both can be considered as factual.			
Question: Which insect gives off the froth known as cuckoo spit?	**Generated Document:** Spittlebugs are sap-feeding insects, members of the Cercopoidea superfamily. They are commonly called spittlebugs or froghoppers because the nymphs produce a spittle-like protective cover around them. The common name also stems from the fact that many spittlebugs jump well and that they are similar to the true (Ranoidea) froghoppers.	**Generated Answer** Spittlebugs	**Gold Answer** Froghopper
Notes: The parent taxon of the generated answer, Spittlebugs, is the gold answer, Froghopper.			

Table 6: Examples where the generated answer has a different surface form than the gold answer, but the factuality filter detects the generated answer as factual.

A Qualitative Examples from Filter

We present qualitative examples from the factuality filter in Table 6 and from the attribution filter in Table 7.

B Prompts

We share the prompts for recitation generation in Figure 5, for factuality filtering in Figure 6, and for attribution filtering in Figure 7.

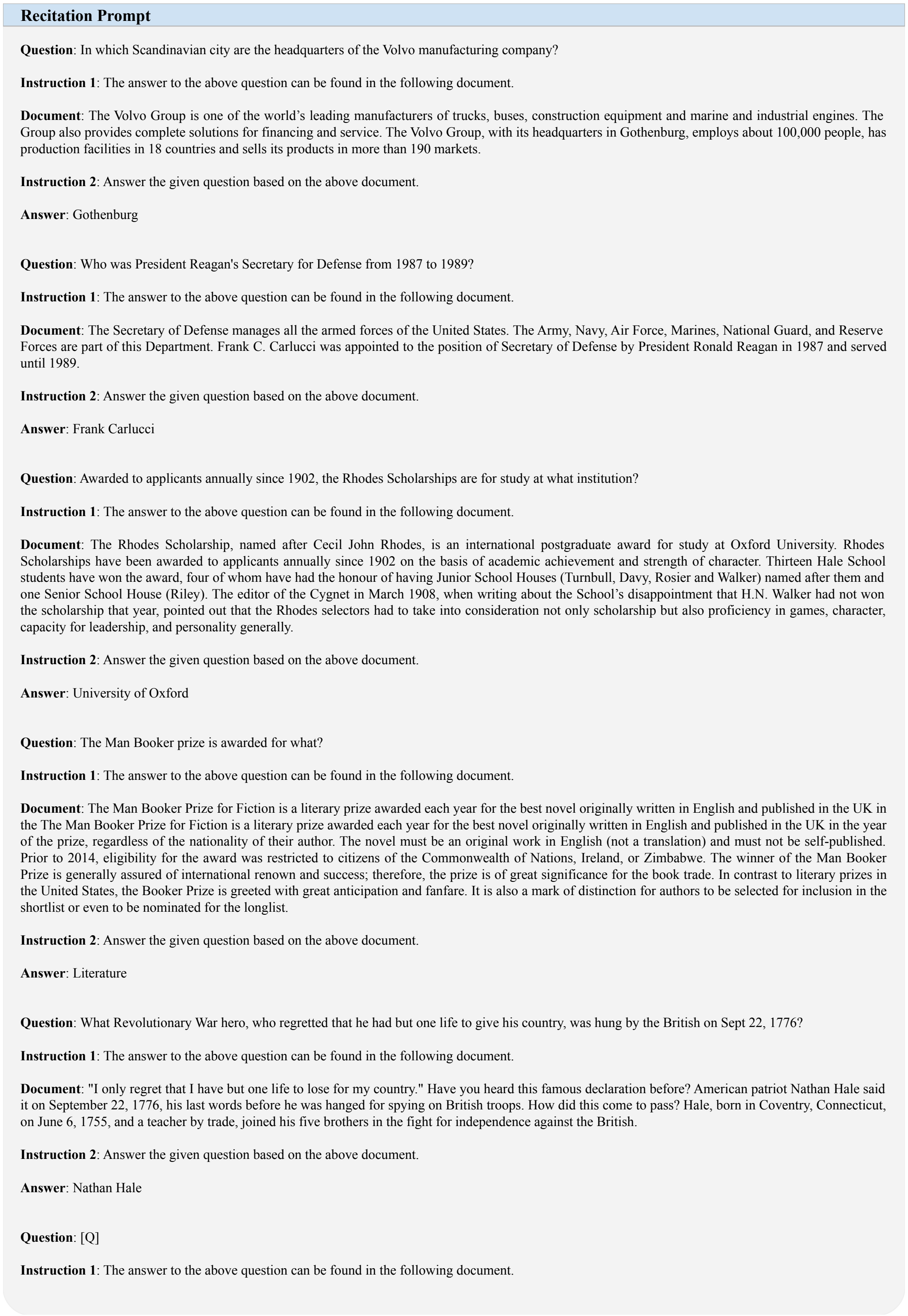
C Additional Results

We share the detailed results with all scores for each out-of-domain dataset for **Q3** in Table 10 and for **Q4** in Table 12.

Since we only share token-level F1 scores in §4, we present the tables with exact match scores. **Q1**: Table 8 and Table 9, **Q2**: Figure 8, **Q3**: Table 11, **Q4**: 13.

Question:	**Generated Document:**	**Generated Answer**	**Gold Answer**
In 2015 which London museum rejected the Thatcher family's offer to display memorabilia belonging to the former Prime Minister?	The museum rejected the Thatcher family's offer to display memorabilia belonging to the former Prime Minister because it was not "in line with the museum's collecting policy". The decision by the museum, which is run by the Imperial War Museums, was made in 2015 but has only now been revealed in a freedom of information (FoI) request made by the Guardian.	The Cabinet War Rooms	The Victoria and Albert
Notes: The counterfactual document mentions the Imperial War Museums, but the generated answer, 'The Cabinet War Rooms', is not grounded in the text, even though it is part of the Imperial War Museums.			
Question:	**Generated Document:**	**Generated Answer**	**Gold Answer**
Which African country has the shilling as it's currency?	The shilling is a unit of currency formerly used in the United Kingdom, Australia, New Zealand and other British Commonwealth countries. The word shilling comes from scilling, an accounting term that dates back to Anglo-Saxon times where it was said to be the value of a cow in Kent or a sheep elsewhere. The shilling was the denomination of the pre-decimal pound sterling, and was worth 12 pence. It was designated by the symbol "s" – not from the first letter of "shilling", but from the Latin solidus. The symbol "s" was used as a shilling abbreviation from the 18th century until decimalisation in 1971.	UK	Kenya
Notes: The counterfactual document does not mention the UK as a part of Africa, therefore the generated answer is not grounded in the document.			
Question:	**Generated Document:**	**Generated Answer**	**Gold Answer**
Who was the first British female singer to have a UK No.1 hit in the 1950's?	Lonnie Donegan (born Anthony James Donegan; 29 April 1931 – 3 November 2002) was a Scottish musician, known as the "King of Skiffle" and a major figure in the British popular music scene in the 1950s and 1960s. In the late 1950s, he was the first British male singer to have two top ten hits in the US Billboard Hot 100 chart and also the first British artist to have a number one US single. He had two number one hits on the British singles chart.	Lonnie Donegan	Lita Roza
Notes: The question asks for the first British *female* singer, but the counterfactual document and the generated answer mention the first *male* singer.			

Table 7: Examples where the attribution filter detects the generated answer as not grounded in the generated document. These are all counterfactual examples, since they fed into the attribution filter after the factuality filter.



Recitation Prompt

Question: In which Scandinavian city are the headquarters of the Volvo manufacturing company?

Instruction 1: The answer to the above question can be found in the following document.

Document: The Volvo Group is one of the world's leading manufacturers of trucks, buses, construction equipment and marine and industrial engines. The Group also provides complete solutions for financing and service. The Volvo Group, with its headquarters in Gothenburg, employs about 100,000 people, has production facilities in 18 countries and sells its products in more than 190 markets.

Instruction 2: Answer the given question based on the above document.

Answer: Gothenburg

Question: Who was President Reagan's Secretary for Defense from 1987 to 1989?

Instruction 1: The answer to the above question can be found in the following document.

Document: The Secretary of Defense manages all the armed forces of the United States. The Army, Navy, Air Force, Marines, National Guard, and Reserve Forces are part of this Department. Frank C. Carlucci was appointed to the position of Secretary of Defense by President Ronald Reagan in 1987 and served until 1989.

Instruction 2: Answer the given question based on the above document.

Answer: Frank Carlucci

Question: Awarded to applicants annually since 1902, the Rhodes Scholarships are for study at what institution?

Instruction 1: The answer to the above question can be found in the following document.

Document: The Rhodes Scholarship, named after Cecil John Rhodes, is an international postgraduate award for study at Oxford University. Rhodes Scholarships have been awarded to applicants annually since 1902 on the basis of academic achievement and strength of character. Thirteen Hale School students have won the award, four of whom have had the honour of having Junior School Houses (Turnbull, Davy, Rosier and Walker) named after them and one Senior School House (Riley). The editor of the Cygnet in March 1908, when writing about the School's disappointment that H.N. Walker had not won the scholarship that year, pointed out that the Rhodes selectors had to take into consideration not only scholarship but also proficiency in games, character, capacity for leadership, and personality generally.

Instruction 2: Answer the given question based on the above document.

Answer: University of Oxford

Question: The Man Booker prize is awarded for what?

Instruction 1: The answer to the above question can be found in the following document.

Document: The Man Booker Prize for Fiction is a literary prize awarded each year for the best novel originally written in English and published in the UK in the The Man Booker Prize for Fiction is a literary prize awarded each year for the best novel originally written in English and published in the UK in the year of the prize, regardless of the nationality of their author. The novel must be an original work in English (not a translation) and must not be self-published. Prior to 2014, eligibility for the award was restricted to citizens of the Commonwealth of Nations, Ireland, or Zimbabwe. The winner of the Man Booker Prize is generally assured of international renown and success; therefore, the prize is of great significance for the book trade. In contrast to literary prizes in the United States, the Booker Prize is greeted with great anticipation and fanfare. It is also a mark of distinction for authors to be selected for inclusion in the shortlist or even to be nominated for the longlist.

Instruction 2: Answer the given question based on the above document.

Answer: Literature

Question: What Revolutionary War hero, who regretted that he had but one life to give his country, was hung by the British on Sept 22, 1776?

Instruction 1: The answer to the above question can be found in the following document.

Document: "I only regret that I have but one life to lose for my country." Have you heard this famous declaration before? American patriot Nathan Hale said it on September 22, 1776, his last words before he was hanged for spying on British troops. How did this come to pass? Hale, born in Coventry, Connecticut, on June 6, 1755, and a teacher by trade, joined his five brothers in the fight for independence against the British.

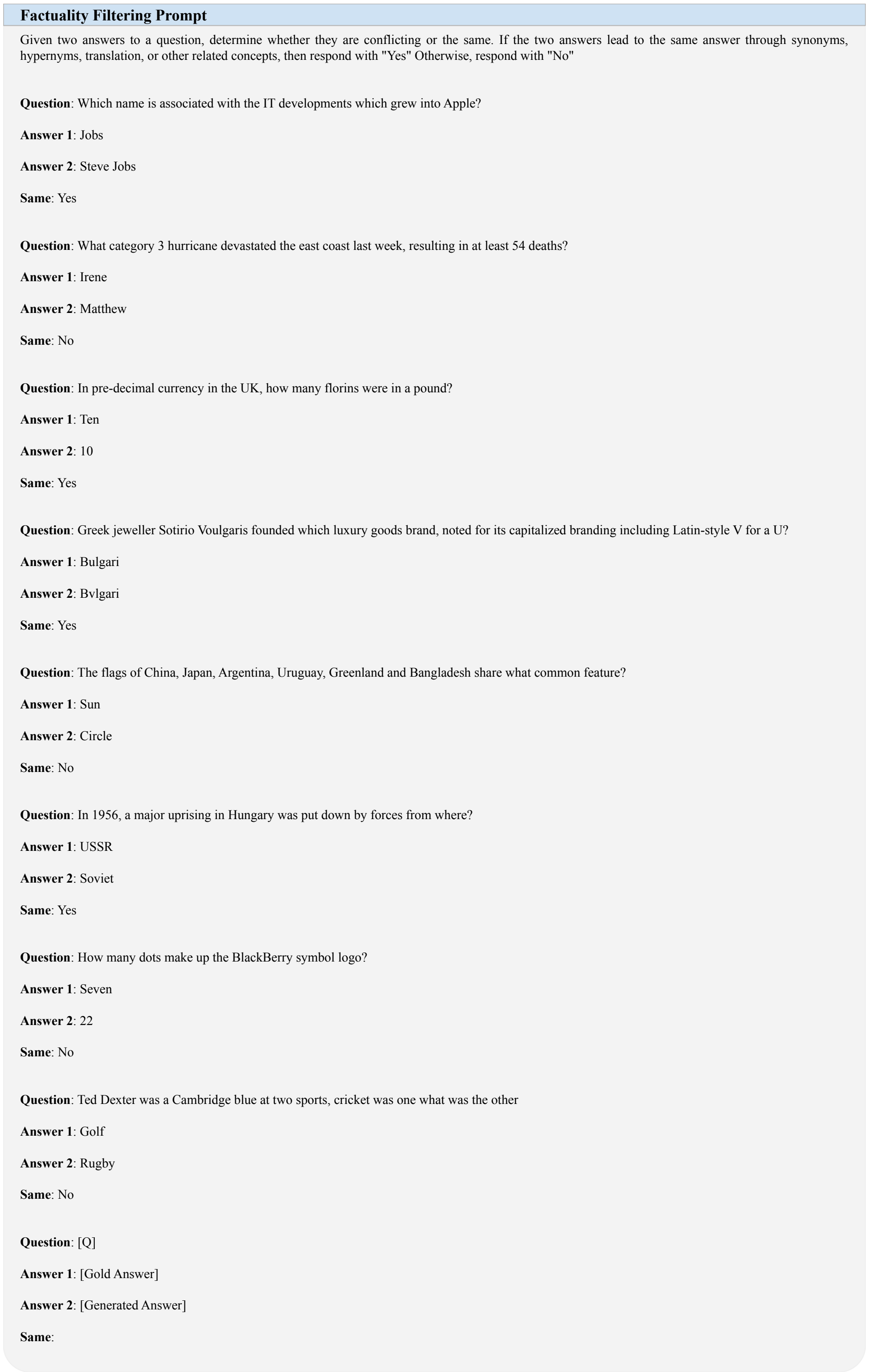
Instruction 2: Answer the given question based on the above document.

Answer: Nathan Hale

Question: [Q]

Instruction 1: The answer to the above question can be found in the following document.

Figure 5: The prompt for the recitation generation step of HAR.



Factuality Filtering Prompt

Given two answers to a question, determine whether they are conflicting or the same. If the two answers lead to the same answer through synonyms, hypernyms, translation, or other related concepts, then respond with "Yes" Otherwise, respond with "No"

Question: Which name is associated with the IT developments which grew into Apple?

Answer 1: Jobs

Answer 2: Steve Jobs

Same: Yes

Question: What category 3 hurricane devastated the east coast last week, resulting in at least 54 deaths?

Answer 1: Irene

Answer 2: Matthew

Same: No

Question: In pre-decimal currency in the UK, how many florins were in a pound?

Answer 1: Ten

Answer 2: 10

Same: Yes

Question: Greek jeweller Sotirio Voulgaris founded which luxury goods brand, noted for its capitalized branding including Latin-style V for a U?

Answer 1: Bulgari

Answer 2: Bvlgari

Same: Yes

Question: The flags of China, Japan, Argentina, Uruguay, Greenland and Bangladesh share what common feature?

Answer 1: Sun

Answer 2: Circle

Same: No

Question: In 1956, a major uprising in Hungary was put down by forces from where?

Answer 1: USSR

Answer 2: Soviet

Same: Yes

Question: How many dots make up the BlackBerry symbol logo?

Answer 1: Seven

Answer 2: 22

Same: No

Question: Ted Dexter was a Cambridge blue at two sports, cricket was one what was the other

Answer 1: Golf

Answer 2: Rugby

Same: No

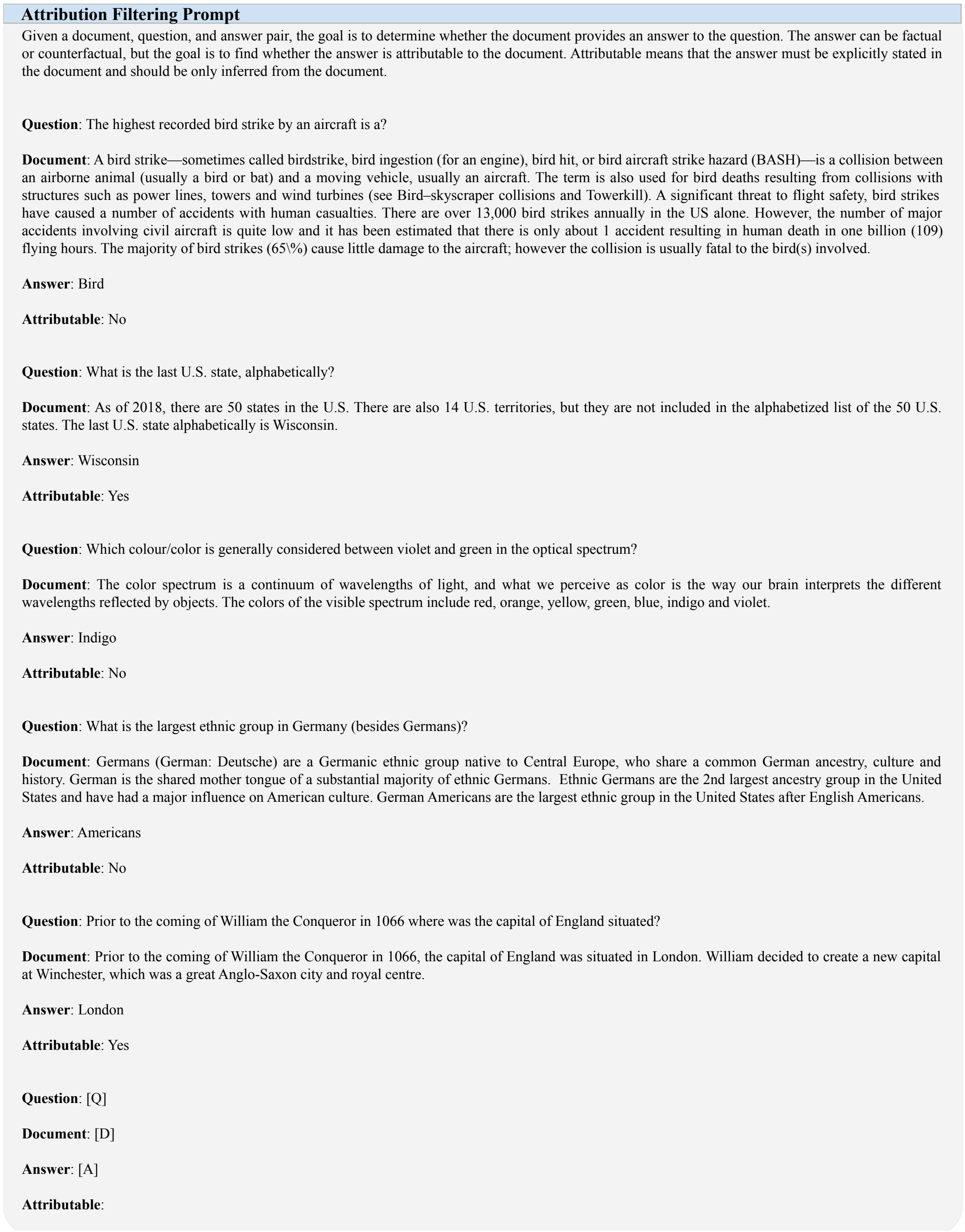
Question: [Q]

Answer 1: [Gold Answer]

Answer 2: [Generated Answer]

Same:

Figure 6: The prompt for the factuality filtering step of HAR.



Attribution Filtering Prompt

Given a document, question, and answer pair, the goal is to determine whether the document provides an answer to the question. The answer can be factual or counterfactual, but the goal is to find whether the answer is attributable to the document. Attributable means that the answer must be explicitly stated in the document and should be only inferred from the document.

Question: The highest recorded bird strike by an aircraft is a?

Document: A bird strike—sometimes called birdstrike, bird ingestion (for an engine), bird hit, or bird aircraft strike hazard (BASH)—is a collision between an airborne animal (usually a bird or bat) and a moving vehicle, usually an aircraft. The term is also used for bird deaths resulting from collisions with structures such as power lines, towers and wind turbines (see Bird–skyscraper collisions and Towerkill). A significant threat to flight safety, bird strikes have caused a number of accidents with human casualties. There are over 13,000 bird strikes annually in the US alone. However, the number of major accidents involving civil aircraft is quite low and it has been estimated that there is only about 1 accident resulting in human death in one billion (109) flying hours. The majority of bird strikes (65\%) cause little damage to the aircraft; however the collision is usually fatal to the bird(s) involved.

Answer: Bird

Attributable: No

Question: What is the last U.S. state, alphabetically?

Document: As of 2018, there are 50 states in the U.S. There are also 14 U.S. territories, but they are not included in the alphabetized list of the 50 U.S. states. The last U.S. state alphabetically is Wisconsin.

Answer: Wisconsin

Attributable: Yes

Question: Which colour/color is generally considered between violet and green in the optical spectrum?

Document: The color spectrum is a continuum of wavelengths of light, and what we perceive as color is the way our brain interprets the different wavelengths reflected by objects. The colors of the visible spectrum include red, orange, yellow, green, blue, indigo and violet.

Answer: Indigo

Attributable: No

Question: What is the largest ethnic group in Germany (besides Germans)?

Document: Germans (German: Deutsche) are a Germanic ethnic group native to Central Europe, who share a common German ancestry, culture and history. German is the shared mother tongue of a substantial majority of ethnic Germans. Ethnic Germans are the 2nd largest ancestry group in the United States and have had a major influence on American culture. German Americans are the largest ethnic group in the United States after English Americans.

Answer: Americans

Attributable: No

Question: Prior to the coming of William the Conqueror in 1066 where was the capital of England situated?

Document: Prior to the coming of William the Conqueror in 1066, the capital of England was situated in London. William decided to create a new capital at Winchester, which was a great Anglo-Saxon city and royal centre.

Answer: London

Attributable: Yes

Question: [Q]

Document: [D]

Answer: [A]

Attributable:

Figure 7: The prompt for the attribution filtering step of HAR. [Q] refers to a question from TriviaQA, [D] and [A] refer to the document and answer pairs generated by LLMs.

Training Dataset	TriviaQA	OOD Avg.
TriviaQA	80.9	50.4
CF-TriviaQA	76.6	**54.0**
TriviaQA +CF-TriviaQA	**81.0**	53.5

Table 8: Exact match scores of T5-3B models finetuned with TriviaQA, CF-TriviaQA, and their combination. Combining our CF-TriviaQA dataset with TriviaQA achieves good out-of-domain performance while having a similar performance in in-domain as the model finetuned with TriviaQA.

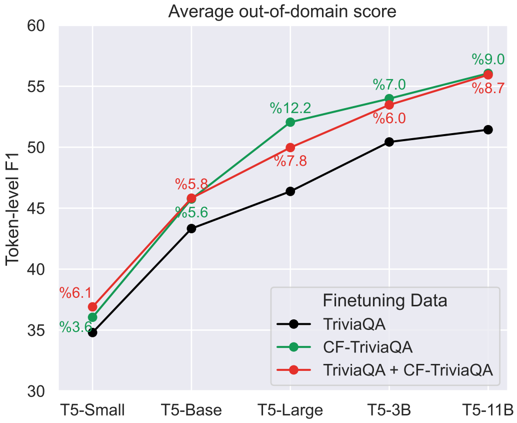


Figure 8: Out-of-domain performance of factual, counterfactual, and combined models with all sizes of T5 models with exact match scores. Models including counterfactual examples consistently outperform factual models across all sizes.

Training Dataset	TriviaQA	SQuAD	NQ	HotpotQA	BioASQ	AQA	AmbigQA	OOD Avg.
TriviaQA	**80.9**	68.6	50.5	51.9	53.3	31.6	46.8	50.4
CF-TriviaQA	76.6	**70.0**	**56.4**	**56.7**	**60.9**	**32.8**	**47.1**	**54.0**

Table 9: Exact match scores of T5-3B models finetuned with TriviaQA vs. CF-TriviaQA. T5-3B model finetuned with CF-TriviaQA significantly outperforms T5-3B with TriviaQA by 3.6 points.

Training Dataset	TriviaQA	SQuAD	NQ	HotpotQA	BioASQ	AQA	AmbigQA	OOD Avg.
TriviaQA	**85.2**	79.6	66.5	69.4	63.4	42.1	53.2	62.4
F-TriviaQA	83.3	80.4	67.7	70.2	**70.2**	41.8	51.3	63.6
CF-TriviaQA	81.7	**81.7**	**71.2**	**73.8**	69.5	**44.9**	**53.2**	**65.7**

Table 10: Token-level F1 scores of T5-3B model finetuned with TriviaQA (human-annotated factual), F-TriviaQA (LLM-generated factual), CF-TriviaQA (LLM-generated counterfactual). While models finetuned with both LLM-generated datasets outperform them model with TriviaQA, the counterfactual model significantly outperforms the factual models, demonstrating the importance of counterfactuals in text grounding.

Training Dataset	**TriviaQA**	**SQuAD**	**NQ**	**HotpotQA**	**BioASQ**	**AQA**	**AmbigQA**	**OOD Avg.**
TriviaQA	**80.9**	68.6	50.5	51.9	53.3	31.6	46.8	50.4
F-TriviaQA	78.6	68.6	52.7	52.3	60.2	30.5	45.7	51.7
CF-TriviaQA	76.6	**70.0**	**56.4**	**56.7**	**60.9**	**32.8**	**47.1**	**54.0**

Table 11: Exact match scores of T5-3B model finetuned with TriviaQA (human-annotated factual), F-TriviaQA (LLM-generated factual), CF-TriviaQA (LLM-generated counterfactual). While models finetuned with both LLM-generated datasets outperform them model with TriviaQA, the counterfactual model significantly outperforms the factual models, demonstrating the importance of counterfactuals in text grounding.

Training Dataset	**TriviaQA**	**SQuAD**	**NQ**	**HotpotQA**	**BioASQ**	**AQA**	**AmbigQA**	**OOD Avg.**
TriviaQA	**85.2**	79.6	66.5	69.4	63.4	42.1	53.2	62.4
CF-TriviaQA$_{\text{PaLM 2-S}}$	80.8	**83.3**	69.7	**73.8**	**71.4**	44.6	**53.2**	**66.0**
CF-TriviaQA$_{\text{PaLM 2-L}}$	81.7	81.7	**71.2**	**73.8**	69.5	**44.9**	**53.2**	65.7

Table 12: Comparison of LLM sizes in the HAR pipeline shows that smaller alternatives of PaLM 2 can achieve similar performance in out-of-domain scores with token-level F1 scores. Therefore, smaller models can be used to utilize hallucination in HAR for better text grounding.

Training Dataset	**TriviaQA**	**SQuAD**	**NQ**	**HotpotQA**	**BioASQ**	**AQA**	**AmbigQA**	**OOD Avg.**
TriviaQA	**80.9**	68.6	50.5	51.9	53.3	31.6	46.8	50.4
CF-TriviaQA$_{\text{PaLM 2-S}}$	75.4	**73.3**	**56.6**	**57.1**	**63.0**	**33.7**	**47.3**	**55.2**
CF-TriviaQA$_{\text{PaLM 2-L}}$	76.6	70.0	56.4	56.7	60.9	32.8	47.1	54.0

Table 13: Comparison of LLM sizes in the HAR pipeline shows that smaller alternatives of PaLM 2 can achieve similar performance in out-of-domain scores with exact match scores. Therefore, smaller models can be used to utilize hallucination in HAR for better text grounding.