Towards Faithful and Robust LLM Specialists for Evidence-Based Question-Answering

Anonymous ACL submission

Abstract

Advances towards more faithful and traceable answers of Large Language Models (LLMs) are crucial for various research and practical endeavors. One avenue in reaching this goal is basing the answers on reliable sources. However, this Evidence-Based QA has proven to work insufficiently with LLMs in terms of citing the correct sources (source quality) and truthfully representing the information within sources (answer attributability). In this work, we systematically investigate how to robustly fine-tune LLMs for better source quality and answer attributability. Specifically, we introduce a data generation pipeline with automated data quality filters, which can synthesize diversified high-quality training and testing data at scale. We further introduce four test sets to benchmark the robustness of fine-tuned specialist models. Extensive evaluation shows that fine-tuning on synthetic data improves performance on both in- and out-of-distribution. Furthermore, we show that data quality, which can be drastically improved by proposed quality filters, matters more than quantity in improving Evidence-Based QA.[1]

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023; Touvron et al., 2023; Anil et al., 2023) have become the center of many cutting-edge applications due to their generalisability and information processing abilities. A typical application of LLMs is *Evidence-Based Question Answering (QA)*, where LLMs are expected to answer questions based on provided sources and cite the sources accurately (e.g., Ni et al., 2023; Vaghefi et al., 2023; Cui et al., 2023; Liu et al., 2023a). By providing these additional sources, multiple shortcomings of standalone LLMs, such as hallucination (Ji et al., 2023) and limited knowledge capacity (Hu et al., 2023), can be addressed, thereby enhancing answer traceability (Gao et al., 2024). However, the performance of existing LLMs on Evidence-Based QA is far from perfect. The SOTA close-sourced LLMs and generative search engines have an unignorable rate of hallucinated answers and false citation (Ni et al., 2023; Liu et al., 2023a). Unfortunately, open-sourced LLMs are even less faithful than the already quality-lacking close-sourced LLMs in Evidence-Based QA (Yue et al., 2023; Gao et al., 2023; also see our evaluation in Section 5.1), although they achieve competitive results on general instruction-following benchmarks (Touvron et al., 2023; Tunstall et al., 2023). We argue that this may prevent practitioners from building Evidence-Based QA (or other RAG) applications in a robust way. Therefore, efficient data creation and fine-tuning methods are urgently needed to improve LLMs' Evidence-Based QA performance in target applications.

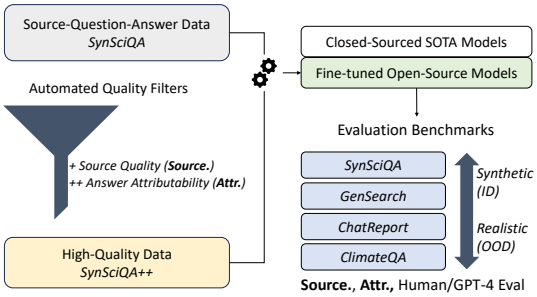

Figure 1: Synthetic data generation pipeline and Evaluation for Evidence-Based QA.

To address this research gap, we first formulate quality dimensions for Evidence-Based QA. Specifically, (1) LLMs need to **always** cite the right evidence at the end of each generated sentence to enable answer traceability, and (2) the answers need to be factually supported by the cited evidence.

Fine-tuning LLMs using Evidence-Based QA data that follow these quality dimensions seems

[1]We will open-source all our codes, LLM generations, and human annotations.

straightforward. However, we identify two major challenges of fine-tuning LLMs into faithful evidence-based question answerers.

C1. Fine-Tuning Data Scalability: Manual annotation for instruction tuning is costly (Conover et al., 2023) and LLM-synthesized data can be a strong alternative (Yin et al., 2023). However, the potentially lower quality of synthesized data may lead to suboptimal fine-tuning performance, given the SOTA LLMs' hallucination rate on Evidence-Based QA (Ni et al., 2023; Liu et al., 2023a).

C2. Generalisability after Fine-tuning: Previous work shows that diversified instruction tuning improves LLMs' generalisability (Chung et al., 2022; Yin et al., 2023). Hence, an intuitive worry is that fine-tuning LLMs (generalists) on Evidence-Based QA data (especially synthetic data) might turn LLMs into specialists that lack generalisability and, thus, struggle with out-of-distribution (OOD) questions and evidence.

To address C1, we propose a data generation pipeline that synthesizes SYNSCIQA (Synthetic Scientific Question Answering), a well-diversified synthetic dataset for Evidence-Based QA. We further extend the pipeline with two novel quality filters to sift out low-quality synthetic data points, leading to SYNSCIQA+ and SYNSCIQA++ (see the left half of Figure 1). To address C2, we first collect an in-domain test set SYNSCIQA$_{test}$ with the data generation pipeline, which shares the data distribution with the training data (i.e., SYNSCIQA) but covers different topics. We further collect three test sets with different distances to the training data distribution to study the OOD performance (see the right half of Figure 1).

Extensive experiments on all proposed train and test settings show that (1) data quality is more important than quantity in Evidence-Based QA fine-tuning; (2) fine-tuning on generated data improves the performance on both in- and out-of-distribution test sets; and (3) performance scores on in-domain test set substantially indicate the OOD performance, suggesting that the synthetic data can be used for validation to estimate the OOD performance. All evaluation metrics are based on golden heuristics and best-performed models from previous work (Yue et al., 2023), which we further verified with human and GPT-4 evaluation. In summary, our contributions include:

1. We propose a data generation pipeline to obtain fine-tuning data for Evidence-Based QA in a salable way, which ensures data diversity and quality.
2. We propose four test sets to benchmark the in- and out-of-distribution performance of fine-tuned Evidence-Based QA specialists.
3. We conduct an extensive evaluation to show that our data-synthesizing strategy leads to effective training and development set for Evidence-Based QA, and quality-filtering significantly improves fine-tuning performance.

2 Evidence-Based Question-Answering

In this section, we formally define Evidence-Based QA. We further define its essential quality dimensions and the corresponding evaluation metrics.

2.1 Task Definition

The task in Evidence-Based QA represents answering a question based on provided sources while truthfully representing and citing the right sources. The model is presented with a set of zero or more *relevant* $\mathcal{S}_{rel}$ and *irrelevant* $\mathcal{S}_{irr}$ sources and a question q. Both are combined in a prompt template $\mathcal{P}$. The model $\mathcal{M}$ is expected to faithfully answer the question and support each answer sentence with a reference to given sources. That is, answer $\mathcal{A} = \mathcal{M}(\mathcal{P}(q, \mathcal{S}_{rel} \cup \mathcal{S}_{irr})) = \{(a_1, s_1); (a_2, s_2); ...; (a_n, s_n)\}$, where n denotes the number of sentences in the answer $\mathcal{A}$; and $s_i \in \mathcal{S}_{cite}$ contains sources cited from $\mathcal{S}_{rel} \cup \mathcal{S}_{irr}$. All answer statements a_i must be attributable to the cited sources s_i rather than the model's parametric knowledge. The only scenario where the model is allowed to answer without citation is when the source evidence doesn't contain question-relevant information. However, the model should address this in its answer. Compared to the answer-attribution task defined in previous work (Li et al., 2023), Evidence-Based QA is more strict as it requires fully attributable and transparent answers.

2.2 Quality Dimensions

We focus on three pivotal quality dimensions to evaluate and improve Evidence-Based QA performance. (1) *Source quality*. This describes whether the model's response only relies on *relevant* sources, and, vice versa, does not include *irrelevant* sources. (2) *Format quality*, i.e., is a citation provided appropriately (to each sentence and in the right format) to maximize the traceability of the information? (3) *Answer attributability*. Given

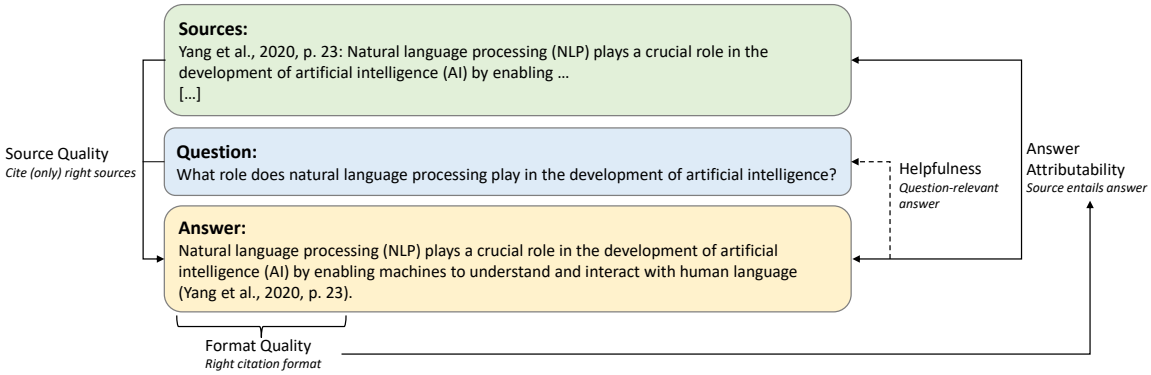

Figure 2: Quality Dimensions of Evidence-Based Question Answering.

correct citation format, an answer sentence is attributable only if it is entailed by the cited source and no hallucination or extrapolation is involved in answering the question. These quality dimensions are reflected in the following prompt template $\mathcal{P}$ which is constantly used in prompting and fine-tuning:

```
Given are the following sources: [BEGIN OF
     SOURCES]
{SOURCE_NAME_1}: {SOURCE_CONTENT_1}
{SOURCE_NAME_2}: {SOURCE_CONTENT_2}
...
{SOURCE_NAME_N}: {SOURCE_CONTENT_N} [END OF
     SOURCES]

Can you respond to the question "{QUESTION}"
     by only relying on the sources. Ignore all
      sources that do not provide an answer to
     the question.
Do not include any knowledge from outside of
     these sources. Only write a single
     paragraph. Each sentence must end with the
      reference in the form of (author, year,
     page number). Strictly follow this format.
      Citing multiple sources in one sentence
     is not allowed.
However, if no source addresses the question,
     admit truthfully that no answer can be
     given.
Answer the question concisely and avoid being
     verbose.
```

By "SOURCE NAME X" and "SOURCE CONTENT X", we denote the X-th source name and content correspondingly. "QUESTION" denotes the question to answer. Note that this prompt is not optimized with prompt engineering tricks. Hence, we hypothesize that our findings can be transferable to practitioners' use cases with different prompt templates.

Besides three quality dimensions, the prompt also requires that more than one citation for one statement is not allowed. We choose this design to maximize the answer traceability and enable clear judgments about attributability by both human and machine evaluators. The NLI models we use are trained on a one-claim-one-evidence setting (Yue et al., 2023) and thus may have suboptimal performance on multi-evidence claim verification, which is more challenging (Jiang et al., 2020).

Our quality dimensions and fine-tuning focus on faithfulness, the most significant shortcoming of open-sourced LLMs in Evidence-Based QA (Yue et al., 2023; Gao et al., 2023). Another important dimension is helpfulness, which can be defined as "how well does the answer address the question?". Our quality dimensions partially address helpfulness by measuring truthful responses based on question-relevant sources. However, we argue that helpfulness is hard to define and evaluate objectively. For this task, it is also challenging to disentangle helpfulness from faithfulness, as a response can only have high helpfulness if it follows the prompt well, i.e., obeying all the faithful citation requirements. To shed light on this aspect, we put additional analyses in Appendix A.

2.3 Evaluation Metrics

We propose two automated metrics using heuristics and automated models to evaluate these quality dimensions in Evidence-Based QA.

Source quality score: Given a prompt $\mathcal{P}(q, \mathcal{S}_{rel} \cup \mathcal{S}_{irr})$ containing zero or more *relevant* sources $\mathcal{S}_{rel}$ and *irrelevant* sources $\mathcal{S}_{irr}$, the model outputs an answer $\mathcal{A}$ citing zero or more sources $\mathcal{S}_{cite}$. Then, the source quality of a sentence is a binary variable described by the following formula:

$$SQ^A = \begin{cases} 1, & \text{if } (|\mathcal{S}_{cite}| > 0) \wedge (\forall s_i \in \mathcal{S}_{cite} : s_i \notin \mathcal{S}_{irr}) \\ 1, & \text{if } (|\mathcal{S}_{cite}| = 0) \wedge (|\mathcal{S}_{rel}| = 0) \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In simple words, source quality equals one if no *irrelevant* source is cited and if a non-zero amount *relevant* sources is given, then the answer must contain a non-zero amount of citations. Otherwise, source quality equals zero.

Attributability score: Given an answer $\mathcal{A}$ with at

least one citation, the attributability score of this instruction-answer pair can be calculated as:

$$Attr.^{\mathcal{A}} = 1 - \frac{|\mathcal{A}_{un}| + |\mathcal{A}_{format}|}{|\mathcal{A}|} = \frac{|\mathcal{A}_{en}|}{|\mathcal{A}|} \quad (2)$$

where $\mathcal{A}_{en}$ ($\mathcal{A}_{un}$) denotes the collection of factually entailed (unentailed) sentences, and $\mathcal{A}_{format}$ denotes the collection of answer sentences with a wrong format or without citation. While the format quality is easy to measure through heuristics, the answer's sentence-source entailment is challenging and requires neural model prediction. In this work, we aggregate the best-performing attribution-prediction models of previous work: attrscore-flan-t5-xl and -xxl checkpoints from Yue et al. (2023) to measure entailment. To achieve higher precision, a sentence is entailed by the cited source only if both models predict "attributable". The attributability score is not applicable for answers without any citation since models should not cite when there is no relevant source. Those answers are addressed by source quality scores (i.e., the model should cite when there is a relevant citation). We mostly follow "citation recall", a metric introduced by Gao et al. (2023), to design attributability scores but adjust it to our stricter setting of Evidence-Based QA (more details in Appendix B).

3 Training Data Generation

Manually annotating Evidence-Based QA data that fulfills all quality dimensions is costly and lacks scalability. In this section, we introduce a novel data generation pipeline to obtain high-quality synthetic data. First, we use OpenAI LLMs to create a diverse and broad base data set of task-specific instruction-answer pairs (SYNSCIQA). Second, we use our quality dimensions to create data sets of higher quality (SYNSCIQA+ and SYNSCIQA++), enabling explorations on the importance of data quality (Zhou et al., 2023).

3.1 SynSciQA

We create SYNSCIQA leveraging both GPT-3.5 and GPT-4 to improve data diversity (GPT-3.5 contributes 75%, see Appendix C for more details). The data creation process proceeds in the following steps:

1. Generate a broad array of 100+ scientific topics.
2. Generate 25 distinctive questions for each topic.
3. Create three source paragraphs *relevant* to each question.
4. Design an instruction encompassing 0-3 *relevant* sources and 3-6 *irrelevant* sources, along with the corresponding question (refer to the prompt template in Section 2.2).
5. Create an answer to the question following the provided instruction.

After the creation process, we split the data into a training set and a test set by topic. This allows us to test different topics in contrast to what we trained the model on and mitigates concerns about data leakage. Using this procedure leaves us with SYNSCIQA comprising 2143 training samples.[2]

3.2 Automated Quality Filters

The clear task definition and data creation process allow us to apply quality filters on the dataset. First, we apply a *source quality* filter to the original SYNSCIQA. Through its construction process, we know which sources are *relevant* and *irrelevant* to the question. Thus, we filter out data points that do not achieve a *source quality* score of one. This leaves us with 1386 samples which we call SYNSCIQA+.

Second, we also apply the *answer attributability* quality dimension as a filter to the dataset. A data point passes the attributability quality filter only if it obtains a full attributability score. This aims to ensure that answers are in the right format and entailed by the sources. Finally, our highest-quality SYNSCIQA++ dataset contains 669 samples.

Since the entailment models in *answer attributability* cannot be controlled with heuristics, we further perform a hand-annotation on 300 randomly-sampled source-answer pairs in the SYNSCIQA++ dataset. Specifically, two annotators per sample investigate whether an answer sentence truthfully reflects the information in the referenced source. We find that both annotators agree on entailment for 94% of the cases, are indecisive for 4%, and conclusive about actual non-entailment for only 2% of the cases (for more details, see Appendix D). These results solidify the validity of the approach.

4 Evaluation Datasets

To assess the validity of the resulting models and their in- and out-of-distribution generalisability, we create a series of four evaluation datasets. The

[2] Some of the API requests were rejected, thus some topics have less than 25 questions.

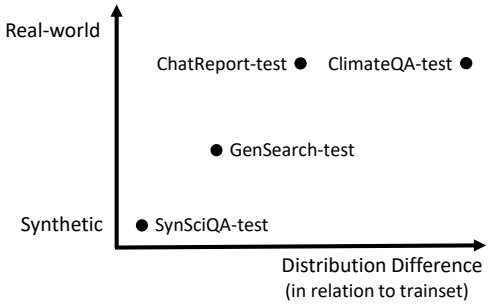

Figure 3: Evaluation Dataset's orientation towards real-world use case scenarios vs. their distribution's proximity to the trainsets.

key differences are whether the data stems from a synthetic or real-world use case and how close the underlying data distribution is to the SynSciQA trainset. Our first evaluation benchmark is SynSciQA$_{test}$ which comprises 539 samples.

Our second evaluation benchmark is GenSearch$_{test}$. This is a dataset adapted from Liu et al. (2023b). In this project, the authors create a dataset from posed questions to generative search engines and mark the question-relevant text part in the given source. We take this dataset and hand-evaluate 600 question-source pairs to distill full-text questions and clear corresponding sources that contain distinct variations of the information. This results in 276 question-source pairs or 106 questions with an average of 2.6 *relevant* sources. After retaining this dataset, we can follow the creation process of SynSciQA (see Step 4 in section 3.1). See Appendix E for more details.

We further create ChatReport$_{test}$. ChatReport is an open-source RAG tool that analyses companies' (sustainability) reports (Ni et al., 2023). It uses eleven sustainability-related questions to analyze the company's disclosure. Inherently, RAG systems' answers rely on source paragraphs from the underlying document, i.e. the company's report. Thus, we use the top-10 most relevant paragraphs (retrieved by ChatReport source code) as input for our system and create 110 instructions. This means we leave the structure of *relevant* / *irrelevant* sources and adopt a genuine RAG setting.

Finally, we use another RAG tool to create ClimateQA$_{test}$. ClimateQA[3] is a RAG system that answers questions based on IPCC and IPBES reports. We pose 261 climate-related questions from Welch (2022) to the system and store the outputted sources. Again, we use this data as input for our instruction form.

Figure 3 illustrates the distance between proposed test sets and SynSciQA$_{test}$ in dimensions of use case and data distribution. ChatReport$_{test}$ and ClimateQA$_{test}$ are directly extracted from real applications while GenSearch$_{test}$ is also from real research engine retrieval but with manual parsing (semi-synthetic). They also have different distribution distances to SynSciQA. GenSearch$_{test}$ contains vastly diversified, non-scientific questions; the source texts of ChatReport$_{test}$ contain formatting noise from sustainability reports; and ClimateQA$_{test}$ contains nested citations in its source texts, which may influence the models' citation correctness. Further explorations for each test set are showcased in Appendix F.

5 Experiments

This section introduces our experiments and analyses in detail. We conduct experiments on Llama-2-chat-13b (Touvron et al., 2023) and Zephyr-7b-β (Tunstall et al., 2023). These models are chosen because they are from two widely used model families: Llama-2 and Mistral (Jiang et al., 2023). Their architecture can be representative of similar causal LLMs. We use aligned models instead of their base models (Llama-2-13b and Mistral-7b) to have models better understand the required quality dimensions for Evidence-Based QA. We use QLoRA (Dettmers et al., 2023) and greedy decoding for all LLM fine-tuning and inference correspondingly. Hyperparameters and other settings are presented in Appendix G.

5.1 Zero-Shot Performance

We first use the proposed test sets and evaluation metrics to benchmark the zero-shot performance of close-sourced and open-sourced LLMs on Evidence-Based QA. The results are shown in Table 1. We find that there is a significant performance gap between open- and close-sourced LLMs on Evidence-Based QA, although they achieve comparable performance on general instruction-following benchmarks (for instance, Zephyr-7b-β vs. GPT-3.5 on MT-Bench (Tunstall et al., 2023)). Similarly, Tunstall et al. (2023) shows that Zephyr-7b-β outperforms Llama-2-70b-chat on all dimensions of MT-Bench, while our evaluation shows that Llama-2-13b-chat hallucinated less than Zephyr-7b-β on Evidence-Based QA. Therefore, the proposed Evidence-Based QA benchmarks can

[3]https://huggingface.co/spaces/Ekimetrics/climate-question-answering

	SynSciQA$_{test}$		GenSearch$_{test}$		ChatReport$_{test}$		ClimateQA$_{test}$	
	Source.	Attr.	Source.	Attr.	Source.	Attr.	Source.	Attr.
Llama-2-13b-chat	49.91	21.56	69.80	7.12	-	8.79	-	2.13
Zephyr-7b-β	36.92	12.90	66.98	3.74	-	5.22	-	2.16
GPT-3.5	53.25	59.65	96.23	40.39	-	34.85	-	18.93
GPT-4	**62.71**	**80.31**	**99.06**	**46.55**	-	**43.37**	-	**27.84**

Table 1: Zero-shot performance of popular open- and close-sourced LLMs on proposed Evidence-Based QA benchmarks. Source. and Attr. are short for source quality, and attributability correspondingly. Source quality is not applicable for ChatReport$_{test}$ and ClimateQA$_{test}$ as source-relevance labels are not available for these real RAG systems (see discussions in Appendix H).

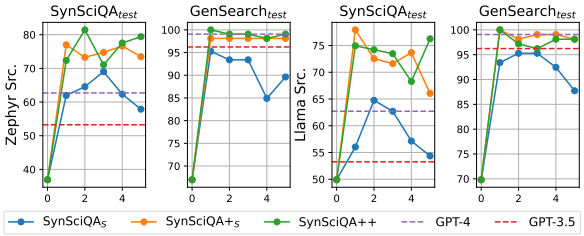

Figure 4: Controlling **quantity**, Source Quality scores vs. number of epoch, caused by different **quality**.

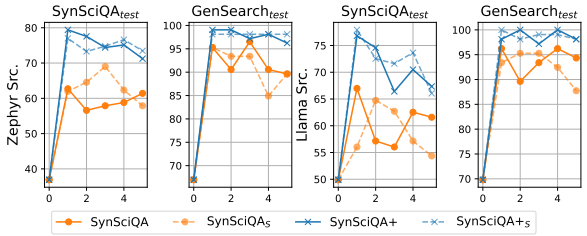

Figure 6: Controlling **quality**, Source Quality scores vs. number of epoch, caused by different **quantity**.

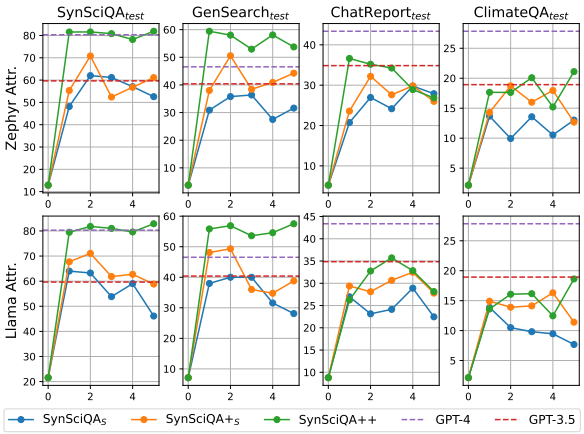

Figure 5: Controlling **quantity**, Attributability scores vs. number of epoch, caused by different **quality**.

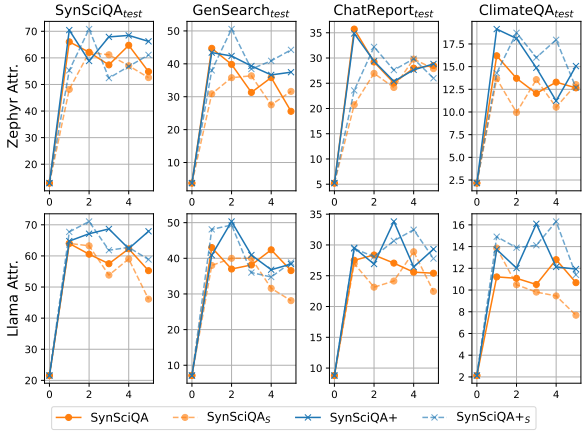

Figure 7: Controlling **quality**, Attributability scores vs. number of epoch, caused by different **quantity**.

be an effective resource to benchmark LLMs' faithfulness, supplementing MT-Bench.

All models achieve lower attributability scores on non-synthetic test sets, indicating that these more realistic settings are more challenging and current LLMs are far from faithful in Evidence-Based QA. The source quality scores on GenSearch$_{test}$ are relatively high since its questions and corresponding sources are extremely diversified (see Appendix E). Thus, it is easier to tell whether a source is relevant to a question or not.

5.2 SynSciQA Fine-Tuning

Given the unsatisfactory performance of open-sourced LLMs on Evidence-Based QA, we want to explore two research questions: **RQ1.** Do data quality and quantity matter for fine-tuning performance with synthetic data? **RQ2.** Can synthetic fine-tuning and evaluation contribute to the performance on OOD data and real-world applications?

To study **RQ1**, we fine-tune open-sourced LLMs on SynSciQA datasets of different qualities. To control quantity when comparing quality, we randomly sample subsets of SynSciQA and SynSciQA+, leading to SynSciQA$_S$ and SynSciQA+$_S$ with the same data quantity as SynSciQA++. To study **RQ2**, we evaluate all fine-tuned checkpoints on test sets of different distributions to see if synthetic data fine-tuning leads to overall improvement. We further calculate the correlation between in-domain (SynSciQA$_{test}$) and OOD (other three test sets) performance to check if in-domain performance can indicate real-world

Comparison	Attr.	Source.
SYN$_S$ < SYN	0.3224	0.5760
SYN+$_S$ < SYN+	0.8719	0.9932
SYN$_S$ < SYN+$_S$	2.88e-3*	7.11e-6**
SYN+$_S$ < SYN++	6.13e-5**	0.1346
SYN$_S$ < SYN++	7.57e-8**	1.02e-5**

Table 2: Statistical significance of performance difference with different trainsets. We conduct Mann-Whitney U test on different settings and use Fisher's method to merge corresponding p-values (see details in Appendix I). By * and **, we denote a p-value smaller than 0.01 and 0.001, respectively. SYN is short for SYNSCIQA

	Zephyr	Llama
SYNSCIQA$_{test}$ & GENSEARCH$_{test}$ Src.	0.99**	0.97**
SYNSCIQA$_{test}$ & GENSEARCH$_{test}$ Attr.	0.96**	0.98**
SYNSCIQA$_{test}$ & CHATREPORT$_{test}$ Attr.	0.94**	0.93**
SYNSCIQA$_{test}$ & CLIMATEQA$_{test}$ Attr.	0.91**	0.94**

Table 3: The performance correlation of different test sets involving all checkpoints of different settings and epochs. The first and second column shows Zephyr-7b-β and Llama-2-13b-chat results correspondingly. By **, we denote a p-value smaller than 0.001.

performance. All fine-tuning lasts 5 epochs and we report the performance of all epochs for two reasons: (1) we suspect that epoch number is an essential hyperparameter for OOD performance, as too many epochs may lead to overfitting to synthetic data; and (2) little previous work explores the influence of epoch number and potential overfitting in instructing tuning.

RQ1: Quality matters more than quantity. We first compare the fine-tuning performance with data of different quality, having the quantity controlled. Figure 4 shows that fine-tuning data with better source quality leads to higher source quality scores (SYNSCIQA+$_S$ and SYNSCIQA++ outperform SYNSCIQA$_S$). Figure 5 shows that higher data quality also leads to better attributability, where in most cases (75%) SYNSCIQA++ > SYNSCIQA+$_S$ > SYNSCIQA$_S$. Fine-tuning on the highest quality data even leads to comparable or better performance than GPT-4 on SYNSCIQA$_{test}$ and GENSEARCH$_{test}$, and GPT-3.5-comparable performance on CHATREPORT$_{test}$ and CLIMATEQA$_{test}$.

Furthermore, when we control quality to compare the fine-tuning outcomes of different quantities, we find that more data points do not lead to significant performance improvement, as illustrated in Figure 6 and Figure 7. We further conduct statistical tests to verify our observations. The results in Table 2 show that improving quality leads to statistically significant improvement while only increasing quantity does not. Therefore, we conclude that data quality is more important than quantity for Evidence-Based QA fine-tuning.

RQ2.1: Fine-tuning on synthetic data positively transfers to real world. It can be observed in Figure 4, Figure 5, Figure 6, and Figure 7 that fine-tuning always lead to better sourcing and attribution performance than original LLMs on in-domain and out-of-distribution test sets. This indicates synthetic data can be used to improve Evidence-Based QA performance in a target domain.

RQ2.2: Synthetic data as validation set for OOD performance. We observe a fluctuating performance corresponding to fine-tuning epochs. Therefore, it is important to conduct checkpoint selection over epochs with a validation (or development) set during fine-tuning. However, performance on SYNSCIQA$_{test}$ is much higher than that of other OOD test sets. Therefore, the performance on a synthetic dataset cannot directly reflect the OOD or real-world performance. But can in-domain synthetic data still be an effective development set indicating which epoch may perform best on OOD data? To answer this question, we compute the Pearson's Correlation between performance scores of different test sets. Results are presented in Table 3, illustrating that the performance on synthetic data has a strong correlation with OOD performance. However, the correlation becomes weaker when the distribution is more distant (CLIMATEQA and CHATREPORT has a weaker correlation than GENSEARCH). Therefore, we conclude that synthetic data can provide a valid development set for OOD performance.

RQ2.3: Overfitting does exist. We suspect that fine-tuning too many epochs may cause overfitting and reduce generalisability. So we compute the Pearson Correlation between the performance scores and epoch numbers of all settings, where positive correlations indicate benefit from more epochs and negative correlations indicate overfits. Results are visualized in Figure 8, showing an overfitting trend for the majority of settings. Therefore, fine-tuning too many epochs may lead to a suboptimal performance. But we do not observe the fine-tuning overfits more to SYNSCIQA than others. We attribute this to SYNSCIQA$_{test}$ containing different scientific topics from the training data.

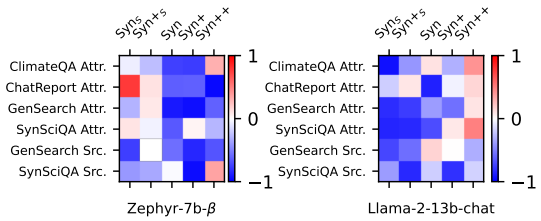

Figure 8: Correlations between performance and epoch number of all settings.

Human vs. Attributability	GPT-4 vs. Attributability	Human vs. GPT-4
0.821**	0.917**	0.871**

Table 4: Pearson Correlation between our Attributability Score, Human Annotation, and GPT-4 Annotation. By **, we denote a p-value smaller than 0.001.

5.3 Validating Attributability Score

Although NLI models have been widely applied in previous work for attributability, they might still be prone to make imperfect predictions (Yue et al., 2023). Therefore, we validate these metrics against human and GPT-4 attributability evaluation. Specifically, we randomly sample instruction-answer pairs from various models and all four evaluation benchmarks and ask humans or GPT-4 to annotate whether the *answer attributability* quality dimension holds for each sentence in the answer.

Then, we calculate the Pearson Correlations between NLI-model-based attributability scores and human / GPT-4 evaluated attributability. As Table 4 shows, the results substantiate the validity of our method for calculating answer attributability. Correlations exceeding 80% across all comparisons between our scores and those annotated by humans and GPT-4 affirm the mutual reinforcement of the outcomes. For more details, see Appendix J.

We also notice a potential shortcut to improve the attributability score: a model may only improve its format quality (i.e., providing a citation to more sentences or more correctly writing source names) without improving the answer entailment rate. In Appendix K, we provide the entailment ratio of format-correct citations as a side result to show that this shortcut does not exist. Both format and attributability dimensions are improved by fine-tuning.

6 Related Work

Basing Answers on Sources: Prompting LLMs to respond with citations has been a popular pattern of Retrieval Augmented Generation (RAG) for better traceability (Karpukhin et al., 2020; Lewis et al., 2021; Borgeaud et al., 2021; Vaghefi et al., 2023; Ni et al., 2023; Asai et al., 2023; Li et al., 2023; Gao et al., 2024). However, previous work shows that asking for citations does not make the answer more factually trustworthy (Min et al., 2023). Commercial search engines and SOTA closed-sourced LLMs suffer from unsatisfactory performance (Ni et al., 2023; Liu et al., 2023a), while open-sourced LLMs have even worse faithfulness (Yue et al., 2023; Gao et al., 2023). Therefore, LLMs, especially open-sourced ones, need essential improvement to achieve more trustworthy RAG applications.

The closest previous study to our work is Gao et al. (2023) which defines evaluation criteria and benchmarks for citation quality of existing LLMs. However, how to scalably fine-tune open-source LLMs in Evidence-Based QA and rigorously evaluate these specialists in- and out-of-distribution remained an open question.

Data Distillation for Instruction Tuning: Distilling instruction-following data from powerful teacher models is an effective and scalable way to improve LLMs' instructing-following performance (Honovich et al., 2023; Wang et al., 2023; Taori et al., 2023; Yin et al., 2023; Tunstall et al., 2023). However, prior research has outlined that simple distillation produces suboptimal data quality (Chen et al., 2023) and that data quality over quantity plays an essential role in improving model output (Zhou et al., 2023). In this work, we propose that automatic filtering can be a potential way to improve distilled data quality and thus achieve better fine-tuning performance.

7 Conclusion

In this work, we present a data synthesize pipeline for fine-tuning and evaluating LLMs for Evidence-Based QA. We show that (1) data quality is critical and our quality filters can effectively improve synthetic data quality; (2) synthetic data fine-tuning can improve real-world RAG use case; and (3) synthetic data can make development set indicating OOD performance. Thus, we advocate the view of specializing and focusing LLMs on specific tasks to reach production-ready, real-world applicable solutions.

Limitations

As with every work, this study has limitations. First, we only experiment on two open-sourced LLMs: Zephyr-7b-β and Llama-2-13b-chat. We hypothesize that the findings are transferable to other pretrained LLMs. We chose this setting because we want to analyze it as comprehensively as possible with our given time and budget restrictions and the broad coverage of investigated aspects including data quantity vs. quantity, out-of-distribution generalisability, and overfitting caused by epoch number.

Second, due to these budget and time limitations, we also conduct random sampling when performing human and GPT-4 evaluation to verify the attributability score instead of evaluating all instruction-answer pairs in all settings and epochs. However, we argue that the performance of different settings is uniform across different instructions and sampled data points are representative enough. Furthermore, we make all hand evaluations and generations of all checkpoints publicly available (for more details, see Appendix J).

Third, this work does not fully assess the quality dimension of helpfulness. We seek to improve open-sourced LLMs in the dimensions of faithfulness and answer-traceability, the most significant shortcomings of open-sourced models. We argue that helpfulness is hard to define and leave it to future exploration (see Appendix A).

Fourth, this work only explores a single prompt template for Evidence-Based QA, which states our quality dimensions and extra requirements for better traceability (e.g., one sentence one citation). Since we conduct no prompt engineering/optimization, we hypothesize that the core findings of this work are transferable to other use cases where prompt templates need to be different (e.g., different citation format, evidence-grounded RAG tasks other than QA). Specifically, practitioners may depend on their own need to write prompt templates and define quality filters to improve distilled data. We plan to verify this in future work.

Ethics Statement

Human Annotation: In this work, all human annotators are Doctorate, Post-Doc researchers or Professors who have good knowledge about scientific communication and entailment. They are officially hired and have full knowledge of the context and utility of the collected data. We adhered strictly to ethical guidelines, respecting the dignity, rights, safety, and well-being of all participants.

Data Privacy or Bias: There are no data privacy issues or bias against certain demographics with regard to the data collected from real-world applications and LLM generations. All artifacts we use are under creative common licence. We also notice no ethical risks associate with this work

Reproducibility Statement: To ensure full reproducibility, we will disclose all codes and data used in this project, as well as the LLM generations, GPT-4 and human annotations. For OpenAI models, we use gpt-3.5-turbo-0613 and gpt-4-0613 for synthetic data generation and gpt-4-turbo-0125-preview for GPT-4 evaluation (due to the project timeline, we do not use gpt-4-turbo-0125-preview for synthetic data generation). We always fix the temperature to 0 when using APIs.

References

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. 2023. Palm 2 technical report.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ArXiv:2310.11511 [cs].

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023. Alpagasus: Training a better alpaca with fewer data.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. ArXiv:2210.11416 [cs].

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. ArXiv:2306.16092 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv:2312.10997 [cs].

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *ArXiv*, abs/2310.05177.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACMComputing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. ArXiv:2004.04906 [cs].

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.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A Survey of Large Language Models Attribution. ArXiv:2311.03731 [cs].

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating Verifiability in Generative Search Engines. ArXiv:2304.09848 [cs].

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023b. Evaluating verifiability in generative search engines.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. ArXiv:2310.16944 [cs].

Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Dominik Stammbach, Nicolas Webersinke, et al. 2023. Chatclimate: Grounding conversational ai in climate science.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.

Ivo Welch. 2022. *What Should You Know About Climate Change?: A Collection of Frequently Asked Questions With Simple Answers For The Concerned*.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4031–4047, Singapore. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.

A The Challenge of Helpfulness

Helpfulness can be defined as "How well does the answer address the question?". We argue that helpfulness is extremely hard to evaluate in Evidence-Based QA.

	SynSciQA$_{test}$			GenSearch$_{test}$			ChatReport$_{test}$			ClimateQA$_{test}$		
	avg. # sentences	avg. length sentences	avg. unique citations	avg. # sentences	avg. length sentences	avg. unique citations	avg. # sentences	avg. length sentences	avg. unique citations	avg. # sentences	avg. length sentences	avg. unique citations
Llama-Syn	4.30	27.01	2.01	2.76	20.82	1.31	4.21	24.44	1.67	5.34	24.24	3.15
Llama-Syn++	3.89	28.06	1.83	2.13	21.26	1.25	3.78	25.73	1.46	4.15	24.71	2.88
Zephyr-Syn	4.37	27.72	2.15	2.62	19.56	1.38	4.58	24.97	1.94	5.20	25.24	3.49
Zephyr-Syn++	3.89	26.50	1.56	2.08	19.04	1.27	3.80	23.60	1.56	4.26	22.73	2.70

Table 5: Statistics of fine-tuned models' outputs. Larger values are underlined. The table reports the second-epoch checkpoints' outcomes.

Following the definition of helpfulness, one could argue that the comparison of question-answer pairs can yield insights into helpfulness. If the model answers the question well, then it is helpful. However, as Table 6 shows, this undermines the logic of answering based on sources in Evidence-Based QA. If no sources are given, then the answer should reflect that. Following the definition, this might be less helpful but certainly more faithful. We argue that this example rather shows that helpfulness and faithfulness are intertwined. Therefore, we view that our source quality score partially addresses helpfulness by indicating whether the answer is only based on valid sources.

Secondly, generations of fine-tuned models are driven by the distribution fine-tuning data. As illustrated in Table 5, models fine-tuned on SynSciQA++ result in slightly shorter answers and a smaller number of unique citations than SynSciQA. This perfectly reflects the training data distribution (see Table 8). Following the definition of helpfulness, one could argue that more context and therefore more answer length is more helpful. However, longer answers with more citations do not indicate more helpfulness in Evidence-Based QA. Table 7 shows an example where one answer sentence - irrespective of source quality - concisely answers the question while the other provides extra context. Is the answer with more context more helpful? We argue that it highly depends and is therefore not easily evaluable. However, through the lens of Evidence-Based QA, an answer is only helpful if the cited sources entail the answer. Thus, also our second metric of answer attributability partially addresses helpfulness. In addition, if more lengthy answers are preferred, one can easily achieve that by encouraging longer answers when generating and filtering synthetic data, as shown by Table 5 and Table 8.

Collectively, we view the exact measurement of helpfulness as a challenge for future work. But we argue that our two employed evaluation metrics already address helpfulness in Evidence-Based QA to a satisfactory degree.

B Attributability Score Details

We use SpaCy (Honnibal and Montani, 2017) to split answers into sentences. Unattributable answer sentences caused by missing citations or wrong citation format can be easily identified through golden heuristics, for example, matching citations with actual source names. However, it is hard for a heuristic-based method to judge whether a statement is entailed by the cited source. Previous work proposes to use NLI models to predict entailment (Honovich et al., 2022; Gao et al., 2023; Yue et al., 2023). Among them, Yue et al. (2023) aggregate the largest NLI training set and conduct extensive analyses to explore the best practice of attributability prediction. Therefore, we rely on their results to select models for the attributability score. The two best-performing checkpoints are Flan-t5-XL and Flan-t5-XXL [4]. When inferencing with these checkpoints, we follow the prompt template in Yue et al. (2023) and use greedy generation. We aggregate the prediction of both models to improve the precision since false positives are more harmful than false negatives in the task of judging LLM faithfulness.

The design of our attributability score mostly follows the citation recall score of Gao et al. (2023). However, we only calculate attributability scores on answers with at least one citation, which differs from Gao et al. (2023), because we also consider scenarios where there is no relevant source at all. In that case, the model should state no source is relevant without any citation.

C Training Data Creation Process

For creating the raw data, we employ several steps. For all creation steps, we use the June checkpoints of GPT-3.5-turbo and GPT-4. In the following, the used prompts are displayed. First, we create a set of 100+ random topics with the help of GPT-4 using the following prompt.

[4] https://huggingface.co/osunlp/attrscore-flan-t5-xxl

Question:	**How does natural selection interact with genetic drift?**
Answer by GPT-4	*None of the provided sources address the interaction between natural selection and genetic drift.*
Answer by Llama-13b-chat	*Sure, I'd be happy to help! Natural selection and genetic drift are both important mechanisms that drive the evolution of species. While natural selection is the process by which individuals with advantageous traits are more likely to survive and reproduce, genetic drift is the random change in the frequency of alleles in a population over time (Morse et al., 2018, p. 67). The interaction between natural selection and genetic drift can be complex, as both processes can influence the evolution of a species in different ways. For example, natural selection can lead to the fixation of advantageous alleles in a population, while genetic drift can result in the loss or fixation of neutral or deleterious alleles (Hudsonsonian Institution, 2017, p. 28). Additionally, [...]*
Answer by Llama-13b-chat fine-tuned for two epochs on SynSciQA++	*None of the provided sources address the question "How does natural selection interact with genetic drift?" Therefore, an answer cannot be given based on these sources.*

Table 6: Answers to the same prompt containing no question-relevant sources by GPT-4, plain vanilla LLama-13b-chat and LLama-13b-chat fine-tuned for two epochs on the high-quality dataset SynSciQA++. Cited relevant sources are marked in green, and irrelevant or erroneous sources are marked in brown.

Question:	**Who is the top ranked female tennis player in the world?**
Answer by GPT-4	*The current top-ranked female tennis player in the world is Iga Swiatek (Online2602022, 2019, p.8).*
Answer by Zephyr-7b-β	*As of the sources provided, Iga Swiatek currently holds the top ranking in women's singles tennis. This information can be found in the source "Online2602022" (Williams, 2022, p.8). Ashleigh Barty, the previous world No.1, retired early in 2022, which led to Swiatek's ascension to the top spot. (Williams, 2022, p.8)*
Answer by Zephyr-7b-β fine-tuned for two epochs on SynSciQA++	*The top ranked female tennis player in the world is Iga Swiatek (Online2602022, 2019, p.8).*

Table 7: Answers to the same prompt containing question-relevant sources by GPT-4, plain vanilla Zephyr-7b-β and Zephyr-7b-β fine-tuned for two epochs on the high-quality dataset SynSciQA++. Cited relevant sources are marked in green, and irrelevant or erroneous sources are marked in brown.

```
"Create {n} random topics from the scientific
     areas of finance, sustainability, physics,
      social sciences and natural sciences.
     Please seperate each topic with '||'. Use
     no enumeration or additional signs to
     seperate the topics."
```

This results in a broad span of topics ranging from *"Corporate finance"* over *"Anthropology"* and *"Electromagnetism"* to *"Dark matter"*. Following this first step, we create 25 questions per topic with GPT-4 (see below).

```
"Take the topic {topic} and create {n}
     questions that could be posed in the field
     . Make the questions diverse and
     differentiable from each other.

End every question with '\\'. Use no
     enumeration or additional signs to
     seperate the questions."
```

Furthermore, we create three paragraphs that address the question as an artificial source with both GPT-3.5 and GPT-4. A random variable is introduced that enforces the creation of around 25% of the data points with GPT-4 to enhance the diversity of the training dataset distribution. The exact final percentage for data created with GPT-4 is 24.97%.

```
"Consider the following question within the
     topic {topic}: {question}

Please create {m} paragraphs with the length
     of 2-4 sentences that partially address
     this question. The question should not
     fully be answered by one paragraph but
     rather helpful content in respect to the
     question should be displayed. Each
     paragraph should be in the style of a book
      or research article.

Furthermore, the paragraphs can display
     different perspectives and should not
     overlap much. The paragaphs should also
     alternate in level of detail and addressed
      readers, i.e., some paragraphs can be
     very scientifc while others would rather
     serve a general public.

It is important that the paragraphs stand for
     themselves. They don't read like one
     article but excerpts from multiple
     articles.

Please be creative with the beginning of the
     paragraphs.

In the end of each paragraph give author, year
      and page in the following format '[author
     , year, page]'. Follow this example: '[
     Mishra et al., 2019, p.54]'.

Make up author, year and page, if you don't
     have this information. Authors can also be
      institutions.

End every paragaph with 'ENDOFPARAGRAPH'. Use
     no enumeration or additional signs to
     seperate the paragraphs. Also do not give
     any further information like "Paragraph 1:
      ...".
```

Finally, we design an instruction that contains 0-3 *relevant* sources that stem from the paragraphs created above, and 3-6 *irrelevant* sources that do

not correspond to the question (for a template, see Prompt Template in Section 2.2). For selecting the *irrelevant* sources, we randomly sample sources from other topics in the dataset. We use GPT-3.5-turbo and GPT-4 to create an answer according to the source creation. This results in the SYNSCIQA dataset.

Finally, we apply the source quality filter to obtain SYNSCIQA+ and the answer attributability filter to obtain SYNSCIQA++. Table 8 shows that the instructions stay comparatively similar throughout the filtering process. For the answers, the number of unique citations and the average sentence number slightly decreases after applying the source quality filter and the attributability quality filter correspondingly, indicating that these filters may effectively filter out answers with problematic citations and unattributable statements. This likely coincides with a higher probability of short paragraphs containing fewer errors. However, both mechanisms don't seem to largely influence answer length and number of cited sources. Rather, the intended behavior of concise answers might be strengthened.

D Hand-Evaluation of the Quality Filters

The hand-evaluation of the SYNSCIQA++ dataset centers around the entailment quality, i.e. whether the answer is entailed by the source. The other two quality dimensions, source, and format, can be controlled automatically, i.e. there are only the right sources in the answers and each sentence ends with a source. To control the entailment quality, we randomly sample 300 source-answer pairs that are evaluated by two annotators. The two annotators per sample stem from four researchers including two doctorate researchers, one post-doctorate researcher, and one professor. As Table 9 shows, the overwhelming amount of answers is correctly entailed by the source.

On the one hand, only 2% of the data is not rightfully entailed. These mainly originate from samples where the model replicated the details in a slightly wrong manner. One example can be seen in Figure 9. The answer states that the main organs of the digestive system are the mouth, esophagus, stomach, small intestine, and large intestine. However, this is only one part of the answer. The main organs also comprise the accessory organs (see Figure 9).

On the other hand, 4% of the cases were split decisions. These predominately originate from different interpretations of nuances in the used language. Disagreements are resolved through debating about specific meanings of nuances until a concensus is achieved. Figure 10 shows an example of this.

This analysis shows the limits of the automatic filters that can deal with a good amount of cases but fail to detect the last bit of small nuances. However, since the vast majority of pairs are valid, the quality filters seem to perform the intended way.

E Creation of GENSEARCH$_{test}$

GENSEARCHENGINES-TEST is developed from the dataset created by Liu et al. (2023b). In this project, the authors create a dataset from generative search engines such as Bing Chat, perplexity AI, or NeevaAI. The task in the project is to hand-evaluate different quality dimensions of the answers. Thus, the annotators are presented with queries of these tools and investigate the given sources. Amongst others, they answer whether the source is accurate in answering the question.

We make use of this dataset and hand-check 600 question-source pairs. While evaluating, we quickly identify that some questions should not be taken into account because they are inconclusive, vague, or impractical for other reasons. For instance, the dataset contains questions like *"tips to win fight at school"* or *"Deep web?"*. Additionally, not all sources were practical or necessary to respond to a question. Some questions contained more than ten sources and others contained duplicates. Thus, we hand-filtered 276 question-source pairs that we deemed relevant. This resulted in 106 unique questions with an average of 2.6 sources. We further processed incomplete questions into a question form. For instance, we added a question mark to each question and added fill words if needed to properly understand the question.

Since we now again have a dataset that contains *relevant* and *irrelevant* sources, we can reiterate the steps used for creating SYNSCIQA (see Step 4 in Section 3.1). This way, we create a dataset that is similar in structure but different in the underlying distribution of the data sources and questions. First, the questions are now rather practical and not scientific anymore. This also translates to the source space that is now rather from websites or online blogs. Furthermore, the sources do not necessarily contain full sentences and are usually written by humans in simple language (assuming that online

Dataset	Instructions				Answers		
	# samples	avg. # srcs	# words/src avg.	# words/src std.	avg. # sentences	avg. length sentences	avg. unqiue citations
SYNSCIQA	2143	6.11	112.08	21.64	4.28	27.22	1.93
SYNSCIQA+	1386	6.27	112.21	21.57	4.20	26.78	1.65
SYNSCIQA++	669	6.09	112.39	21.41	3.87	28.49	1.62

Table 8: Overview of the statistical characteristics of the training datasets.

```
SOURCE:  The digestive system is a fascinating and intricate system that allows our bodies to obtain the
     necessary nutrients for survival. While many organs play a role in this process, the main organs of the
      digestive system can be categorized into two groups: the gastrointestinal tract and the accessory
     organs. The gastrointestinal tract consists of the mouth, esophagus, stomach, small intestine, and
     large intestine, which work together to break down food and absorb nutrients. The accessory organs, on
     the other hand, include the liver, gallbladder, and pancreas, which produce and release substances that
      aid in digestion. By understanding the main organs of the digestive system, we can appreciate the
     complexity of this system and the importance of maintaining its health.

ANSWER SENTENCE: [...] The main organs of the digestive system are the mouth, esophagus, stomach, small
     intestine, and large intestine. [...]

HUMAN VERDICT: Two times incorrect entailment. The main organs of the digestive system also comprise the
     accessory organs.
```

Figure 9: Example for an incorrect entailment.

	Percentage
Both annotators agree on correct entailment	94.3%
Both annotators agree on incorrect entailment	2.0%
Split decisions	3.7%

Table 9: Hand-Evaluation Results on Entailment Quality.

Dataset	# samples	avg. # srcs	# words per src avg.	# words per src std.
SYNSCIQA$_{test}$	539	6.10	98.23	21.28
GENSEARCH$_{test}$	106	5.85	82.31	41.80
CHATREPORT$_{test}$	110	10.00	67.94	25.27
CLIMATEQA$_{test}$	261	4.47	134.51	13.35

Table 10: Number of instructions per dataset, average sources per instruction as well as average and standard deviation of number of words per instruction for the test datasets.

articles are written by humans). Since the topic range is much more diverse, the resulting instructions usually contain sources where a human evaluator could clearly state which sources belong to the question. This combination of simple language and a very wide range of questions theoretically makes the differentiation between *relevant* and *irrelevant* sources much easier.

F Test Data Examples

To outline the differences between the test datasets, we further explore the properties of each evaluation benchmark.

The first evident difference lies in the statistical properties of the sources in the instructions (see Table 10). We create SYNSCIQA$_{test}$ and GENSEARCH$_{test}$ with the data creation pipeline described in Section 2.2. This means, we have known *relevant* and *irrelevant* sources in the datasets. On the other hand, we use top-10 retrieved sources for CHATREPORT$_{test}$ and the output sources by ClimateQA for CLIMATEQA$_{test}$. As Table 10 shows, the different sourcing mechanisms result in a large difference in average length and standard deviation.

Furthermore, there are large differences in the structure and format of the sources. The exemplary comparison in Table 11 reveals that SYNSCIQA$_{test}$ and GENSEARCH$_{test}$ are predominantly in full-sentence form. While SYNSCIQA$_{test}$ only contains synthetic scientific topics, GENSEARCH$_{test}$ is created from internet sources and therefore much broader and more colloquial in tone. Sources in CHATREPORT$_{test}$ start in the middle of the sentence, end in the middle of the sentence, and are not necessarily in full-sentence form. The same holds true for CLIMATEQA$_{test}$. However, this dataset also contains nested citations which represents the most complicated case for Evidence-Based QA.

G Hyperparameter and Other Settings

We always use random seed 42 for experiments in this work. We use the default QLoRA hyperparam-

```
SOURCE: From a more technical perspective, working capital management involves the optimization of a firm's
     liquidity position by managing the trade-offs between profitability and risk. This is achieved by
     managing the components of working capital, namely accounts receivable, inventory, and accounts payable
     . The management of these components involves determining the optimal level of investment in each,
     considering the costs and benefits associated with different levels of investment. For instance, while
     a high level of inventory may reduce the risk of stock-outs, it also ties up funds that could be used
     elsewhere in the business.

ANSWER SENTENCE: [...] This is achieved by managing the components of working capital, which include
     accounts receivable, inventory, and accounts payable. [...]

HUMAN VERDICT: One time incorrect entailment. The word "include" means that the following objects are
     necessary but not sufficient while "namely" in the source signals that they are sufficient and
     necessary.
```

Figure 10: Example for an incorrect entailment.

Dataset	Example Source Paragraph
SynSciQA$_{test}$	*The electromagnetic force is responsible for the most familiar interactions in our everyday lives. It is the force that allows us to see, feel, and interact with the world around us. When light interacts with matter, it can be absorbed, reflected, or transmitted, giving rise to the colors we perceive. The electromagnetic force also enables the operation of electronic devices, such as computers and smartphones, by allowing the flow of electric currents. Moreover, this force is essential for generating and transmitting electrical power. Understanding how particles interact through the electromagnetic force is not only of interest to scientists but also has practical applications that impact our daily lives.*
GenSearch$_{test}$	*But how?! Well, there's merchandising, VOD, streaming video, foreign sales, and a plethora of other distribution channels that can help filmmakers, producers, and studios turn a profit. Traditionally, movies have made their money from ticket sales at the box office or in theaters. A studio might make about 60% of a film's ticket sales in the United States, and around 20% to 40% of that on overseas ticket sales.*
ChatReport$_{test}$	*the details of Fortive's climate-related governance, strategy, risk management, and metrics and targets. In 2021, we formally expanded the risk criteria within our Enterprise Risk Management program to account for the financial, operational, and regulatory risks in addition to physical risks for which we were already accounting. By incorporating additional climate-related risks into our existing protocol for evaluating and identifying risk, we are able to capture climate-related*
ClimateQA$_{test}$	*5 that are subducted over decades are expected to experience significant warming (see Figure 5.3). The warming in the subtropical gyres penetrates deeper into the ocean than other gyres (roughly 15ºN–45ºN and 15ºS–45ºS in Figure 5.3), following the wind-driven bowing down of the density surfaces (the solid lines in Figure 5.3) in these gyres (Terada and Minobe, 2018). The greater warming at 700–2000 m in the Atlantic than the Pacific or Indian Oceans (Figure 5.3) reflects the strong southward transport of recently formed NADW at these depths by the AMOC. Two areas that commonly exhibit substantially reduced near-surface warming over the course of the 21st century are the northern north Atlantic, where a slowing AMOC (see Section 6.7.1.1) reduces the northward heat transport and brings the surface temperatures closer to what is found in other ocean basins at these latitudes (Collins et al. 2013), and the southern side of the Southern Ocean, where water upwells*

Table 11: Examples of source paragraphs for the different (training/testing) datasets. The structure of SynSciQA$_{test}$ is representative of SynSciQA.

eter settings [5], namely, an effective batch size of 32, a lora r of 64, a lora alpha of 16, a warmup ratio of 0.03, a constant learning rate scheduler, a learning rate of 0.0002, an Adam beta2 of 0.999, a max gradient norm of 0.3, a LoRA dropout of 0.1, 0 weight decay, a source max length of 2048, and a target max length of 512. We use LoRA module on all linear layers.

We always use SpaCy for word count and sentence split, and Scipy to compute Pearson's Correlation and other statistical significance tests.

All experiments are conducted on two clusters, one with 4 V100 GPUs and the other with 4 A100 (80G) GPUs. 1 GPU hour is used per fine-tuning.

[5] https://github.com/jondurbin/qlora

H Relevance Label for Real RAG

Section 4 introduces that source-relevance label is available for SynSciQA$_{test}$ thanks to the data creation process. We also annotate source-relevance for GenSearch$_{test}$. However, we do not annotate that for ChatReport$_{test}$ and ClimateQA$_{test}$ because we find most of the retrieved top-k sources in those real RAG systems are directly or indirectly relevant since they are retrieved from a narrow domain (e.g., a sustainability report). Take the following source-question pair as an example:

- Question: How resilient is the organisation's strategy when considering different climate-related scenarios, including a 2°C target or lower scenario? How resilient is the organisation's strategy when considering climate physical risks?

- Source: ... Risk Management a. Describe the organization's processes for identifying and assessing climate-related risks. CDP C2.1 CDP C2.2 CDP C2.2a Risk Management b. Describe the organization's processes for managing climate-related risks. CDP C2.1 CDP C2.2 Risk Management c. Describe how processes for identifying, assessing, and managing climate-related risks are integrated into the organization's overall risk management. CDP C2.1 CDP C2.2 Metrics and Targets ...

Although the source does not directly address the resilience of the company's strategy considering climate risks, it provides information about the company's climate-related risk management, which can be indirectly useful for the resilience considering climate-related risk. Therefore, we rely on answer attributability to evaluate the real-world RAG test sets. As long as the answers have good traceability, we assume relevant information is provided to the question.

I Statistical Significance Tests

To show the statistical significance of performance difference in Section 5.2, we first conduct Mann-Whitney U test on each sub-figure of Figure 4, Figure 5, Figure 6, and Figure 7. Specifically, we regard the scores of epoch 1 to 5 comes from the same distribution, and compute if distributions of different settings are statistically significantly different or not. For example, SynSciQA++ distribtuion ([81.56, 81.59, 80.83, 78.19, 81.9]) and SynSciQA$_S$ distribtuion ([48.15, 62.01, 61.17, 57.05, 52.57]) in the first sub-figure of Figure 5. We use Mann-Whitney U test instead of student-t test to avoid making the normal distribution assumption. After having p-values between all settings, we apply Fisher's method to aggregate the p-values, resulting in Table 2.

J Validating the Attributability Score

The target of this validation is to reinforce the validity of the employed methodology in the *answer attributability*. Generally, this investigation follows the same structure as our train set validation in Appendix D. The main difference is that we now investigate all settings, including those out-of-distribution and using open-source models to find evidence that the comparisons are valid. To investigate the decisions, we repeat the evaluation of our score with both human and GPT-4 annotation. Both evaluations follow the structure of *answer attributability* and are articulated through the following prompt.

```
"Your task is to evaluate whether a SENTENCE
     represents the information in a SOURCE.
     This criterion is defined as faithfulness.
      Faithfulness answers the main question of
      "Is the SENTENCE content justified
     through the SOURCE?". The SENTENCE should
     reflect the information given in the
     SOURCE. If the SOURCE information does not
      entail the SENTENCE, then the SENTENCE is
      not faithful. The SENTENCE must not
     contain completely new details that are
     not mentioned in the SOURCE. However, if
     the SENTENCE contains the same meaning as
     the SOURCE but only the wording changes,
     the SENTENCE is still faithful.

SOURCE: +++ {0} +++

SENTENCE: ||| {1} |||

Answer whether the ANSWER is faithful with
     respect to the SOURCE given the above
     definition of faithfulness. Respond by
     starting with "[[YES]]" or "[[NO]]" and
     then justify your decision in at most one
     sentence.
"
```

For human evaluation, we evaluate 8 settings in total: raw models include GPT-3.5, GPT-4, Llama-2-13b-chat, and Zephyr-7b-β; fine-tuned models include Llama-2-13b-chat and Zephyr-7b-β trained on SynSciQA and SynSciQA++ for 2 epochs. We choose the second epoch since it usually does not associate with strong over- or under-fitting. For each setting, we randomly sample 10 instruction-answer pairs from all 4 test sets. Therefore, we evaluate 320 (8 x 4 x 10) datapoints in total for human evaluation. We do random sampling instead of evaluating all settings because hand evaluation of attributability is very costly and time-consuming (for examples of a source-sentence pair in the hand-evaluation, see Figure 9 or Figure 10). In addition, the LLMs have uniform performance on different instructions. We also make all hand evaluations and LLM generations publicly available to justify our hand evaluation choice. Each sample is evaluated by one doctorate researcher. Given the extremely high overlaps in judgments in Appendix D as well as the effort in manual annotation, we choose one annotator per sample to broaden the assessment spectrum.

GPT-4 evaluation is much less expensive than hand evaluation. We thus sample from all 14 raw models and fine-tuning settings. We also sample 25 instruction-answer pairs for each test set. Therefore, we evaluate (14 x 4 x 25) datapoints in total with GPT-4. Table 12 shows all settings in which we conduct human and GPT-4 evaluation.

	GPT-3.5-turbo	GPT-4	Llama-2-13b-chat	Zephyr-7b-β
No Fine-Tuning	Hum./GPT	Hum./GPT	Hum./GPT	Hum./GPT
SynSciQA	-	-	Hum./GPT	Hum./GPT
SynSciQA$_S$	-	-	GPT	GPT
SynSciQA+	-	-	GPT	GPT
SynSciQA+$_S$	-	-	GPT	GPT
SynSciQA++	-	-	Hum./GPT	Hum./GPT

Hum. = Human

Table 12: Evaluation Settings for Annotating Instruction-Answers Pairs with Human (n=10 per setting) and GPT-4 (n=25 per setting) Annotation.

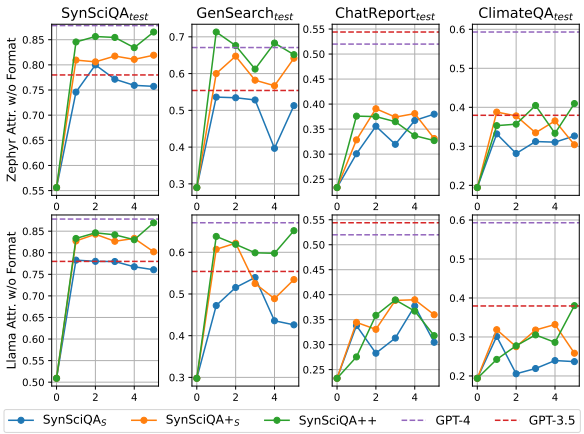

Figure 11: Controlling **quantity**, Attributability scores (without format-wrong sentences) vs. number of epoch, caused by different **quality**.

Finally, we aggregate all available scores to calculate Pearson correlations. For example, we aggregate 32 scores (8 settings x 4 test set) to compute the correlation between human and attributability scores. As Table 4 shows, our *answer attributability* score, the human and GPT-4 annotation arise at majorly the same results. This is signaled by correlation coefficients of over 80%.

K Format Short-Cut in Attributability

The fine-tuning may only improve format quality as a short-cut to improving attributability scores, instead of making the answer sentences more citation-compliant. To verify this, we compute the attributability score again on those format-correct sentences only. In other words, all improvements should then be caused by more sentences supported by sources. The results are shown in Figure 11. It can be observed that in all settings, fine-tuning improves attributability without considering format quality. Interestingly, GPT-3.5 outperforms GPT-4 on ChatReport$_{test}$ ignoring format quality, which coheres to findings in Ni et al. (2023) where GPT-3.5 is better entailed in ChatReport$_{test}$.