On Synthesizing Data for Context Attribution in Question Answering

Anonymous ACL submission

Abstract

Question Answering (QA) accounts for a significant portion of LLM usage "in the wild". However, LLMs sometimes produce false or misleading responses, also known as "hallucinations". Therefore, grounding the generated answers in contextually provided information—i.e., providing evidence for the generated text—is paramount for LLMs' trustworthiness. Providing this information is the task of *context attribution*. In this paper, we systematically study LLM-based approaches for this task, namely we investigate (i) zero-shot inference, (ii) LLM ensembling, and (iii) fine-tuning of small LMs on synthetic data generated by larger LLMs. Our key contribution is SynQA: a novel generative strategy for synthesizing context attribution data. Given selected context sentences, an LLM generates QA pairs that are supported by these sentences. This leverages LLMs' natural strengths in text generation while ensuring clear attribution paths in the synthetic training data. We show that the attribution data synthesized via SynQA is highly effective for fine-tuning small LMs for context attribution in different QA tasks and domains. Finally, with a user study, we validate the usefulness of small LMs (fine-tuned on synthetic data from SynQA) in context attribution for QA.

1 Introduction

Large Language Models (LLMs) have become ubiquitous, with Question Answering (QA) as their most common use case (Trippas et al., 2024). However, LLMs have a tendency to hallucinate: they generate content that is factually incorrect w.r.t. a previously provided reference text. This poses the need for *context attribution* methods that create links between the answer and different relevant parts of the (potentially large) reference text; for an illustration of the task, see Figure 1.

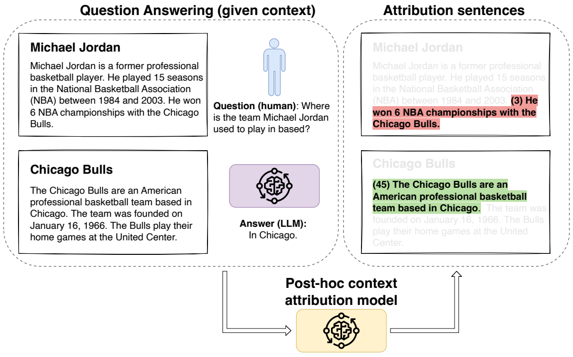

Question Answering (given context)
Attribution sentences
Michael Jordan
Michael Jordan is a former professional basketball player. He played 15 seasons in the National Basketball Association (NBA) between 1984 and 2003. He won 6 NBA championships with the Chicago Bulls.
Question (human): Where is the team Michael Jordan used to play in based?
Chicago Bulls
The Chicago Bulls are an American professional basketball team based in Chicago. The team was founded on January 16, 1966. The Bulls play their home games at the United Center.
Answer (LLM): In Chicago.
Michael Jordan
Michael Jordan is a former professional basketball player. He played 15 seasons in the National Basketball Association (NBA) between 1984 and 2003. (3) He won 6 NBA championships with the Chicago Bulls.
Chicago Bulls
(45) The Chicago Bulls are an American professional basketball team based in Chicago. The team was founded on January 16, 1966. The Bulls play their home games at the United Center.
Post-hoc context attribution model

Figure 1: Post-hoc context attribution: Given a question, an LLM-generated answer, and context (from human input or retrieval), the model identifies supporting sentences within the context. Our user study (§3.3.5) shows that presenting these supporting sentences helps users verify LLM answers more quickly and accurately.

For these reasons, reliable and efficient context attribution is instrumental in manually verifying the factuality of LLM-generated content. By conducting a user study, Slobodkin et al. (2024) report two important findings in this respect: (1) attribution models reduce the human workload in a fact-checking task by as much as 50%; and (2) sentence level granularity—i.e., grounding the answers in one or more relevant sentences in the reference context—is the most efficient granularity level for manual fact-checking of LLM-generated answers.

Given the task importance, recent context-attribution research spans text summarization (Krishna et al., 2023; Ernst et al., 2024), citation attribution (Gao et al., 2023b; Huang et al., 2024b), and question answering (Phukan et al., 2024; Cohen-Wang et al., 2024). However, the solutions rely on document- or paragraph-level evidence, which comes with the following limitations: (1) the user still has to read the (potentially long) document(s) to verify the generated text; and (2) the LLM needs to correctly generate the reference output alongside answering the question correctly. In contrast, the post-hoc attribution methods perform the attribution *after* the LLM generates the answer. These models, however, either attribute to coarse-grained units of text (Nakano et al., 2021; Menick et al., 2022; Buchmann et al., 2024), or provide fine-grained attributions but their inference is computationally expensive (Cohen-Wang et al., 2024), hindering their adoption in practice.

In this paper, we explore how LLMs can generate synthetic data for attribution fine-tuning, enabling accurate, sentence-level, and real-time efficient models. For data generation, we compare two approaches: (1) in the fairly straightforward *attribution synthesis* (SYN-ATT), we start with question-answer pairs from a reference text and prompt the LLM to identify the supporting sentences; (2) in our novel *question-answer synthesis* (SYNQA), we use Wikipedia sentences and prompt the LLM to generate a question-answer pair that is fully supported by these sentences. Given the generated data, we fine-tune smaller, more efficient context attribution models and compare their performance.

Through extensive evaluation encompassing six datasets and two real-world scenarios (attribution for single-turn questions: i.e., a single question and single answer, and for dialogue questions: i.e., as part of a conversation), we demonstrate that models trained on synthetic data generated by SYNQA: ① Outperform zero-shot LLMs that are orders of magnitude larger, while maintaining real-time inference capabilities (§3.3.1); ② Achieve competitive performance on in-domain tasks and superior generalization to out-of-domain datasets compared to models trained on gold data (§3.3.2); ③ Successfully handle dialogue-based attribution without requiring in-domain training data (§3.3.3); ④ Show consistent performance improvements as synthetic training data increases (§3.3.4); ⑤ Significantly improve users' speed and accuracy in verifying LLM-generated answers (§3.3.5). These results highlight the viability of scalable, data-efficient context attribution techniques, thus paving the way for more interpretable and trustworthy AI systems.

2 Synthesizing Attribution Data

Context attribution identifies which parts of a reference text support a given question-answer pair (Rashkin et al., 2023). Formally, given a question q, its answer a, and a context text c consisting of sentences $s_1, ..., s_n$, the task is to identify the subset of sentences $S \subseteq c$ that fully support the answer a to question q. To train efficient attribution models without requiring expensive human annotations, we explore synthetic data generation approaches using LLMs. We implement two methods for synthetic data generation. Our baseline method (SYN-ATT) is discriminative: given existing question-answer pairs and their context, an LLM identifies supporting sentences, which are then used to train a smaller attribution model. Our proposed method (SYNQA) takes a generative approach: given selected context sentences, an LLM generates question-answer pairs that are fully supported by these sentences. This approach better leverages LLMs' natural strengths in text generation while ensuring clear attribution paths in the synthetic training data.

2.1 SYNQA: Generative Synthetic Data Generation Method

SYNQA consists of three parts: context selection, QA generation, and distractors mining (for an illustration of the method, see Figure 2). In what follows, we describe each part in detail.

Context Collection. We use Wikipedia as our data source, as each article consists of sentences about a coherent and connected topic, with two collection strategies. In the first, we select individual Wikipedia articles for dialogue-centric generation and use their sentences as context. In the second, for multi-hop reasoning, we identify sentences containing Wikipedia links and follow these links to create "hops" between articles, limiting to a maximum of two paths to maintain semantic coherence, while enabling more complex reasoning patterns (for more details, see Appendix B).

Question-Answer Generation. Given the set of contexts, an LLM can now generate question-answer pairs. For single articles, we prompt the model to generate multiple question-answer pairs, each grounded in specific sentences. This creates a set of dialogue-centric samples where questions build upon the previous context. For linked articles, we prompt the model to generate questions that necessitate connecting information across the articles, encouraging multi-hop reasoning. This yields multi-hop samples requiring integration of information across documents, as well as samples that mimic a dialogue about a specific topic given the context. We provide the full prompts used for generation in Appendix C.

Distractors Mining. To make the attribution task more realistic, we augment each sample with distractor articles. With E5 (Wang et al., 2022), we embed each Wikipedia article in our collection. For

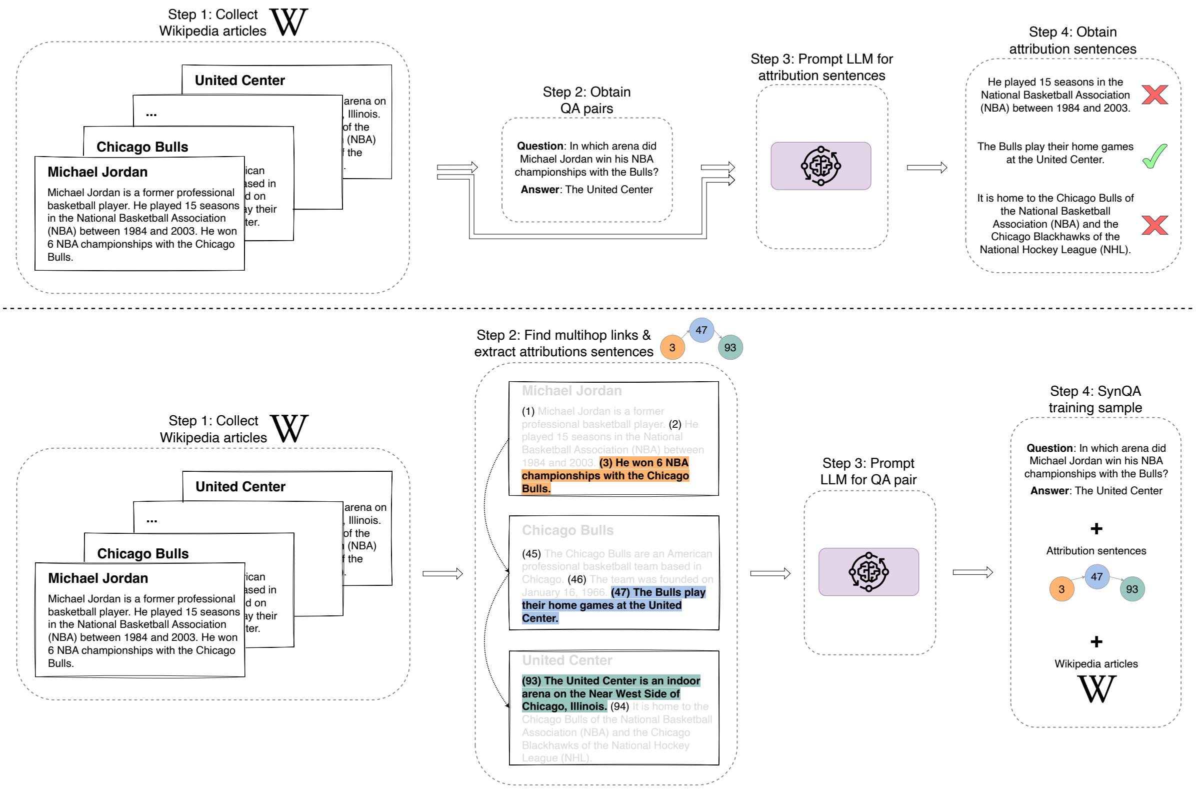

Step 1: Collect Wikipedia articles
W
United Center
...
Chicago Bulls
Michael Jordan
Michael Jordan is a former professional basketball player. He played 15 seasons in the National Basketball Association (NBA) between 1984 and 2003. He won 6 NBA championships with the Chicago Bulls.
Step 2: Obtain QA pairs
Question: In which arena did Michael Jordan win his NBA championships with the Bulls?
Answer: The United Center
Step 3: Prompt LLM for attribution sentences
Step 4: Obtain attribution sentences
He played 15 seasons in the National Basketball Association (NBA) between 1984 and 2003.
The Bulls play their home games at the United Center.
It is home to the Chicago Bulls of the National Basketball Association (NBA) and the Chicago Blackhawks of the National Hockey League (NHL).
Step 1: Collect Wikipedia articles
W
United Center
...
Chicago Bulls
Michael Jordan
Michael Jordan is a former professional basketball player. He played 15 seasons in the National Basketball Association (NBA) between 1984 and 2003. He won 6 NBA championships with the Chicago Bulls.
Step 2: Find multihop links & extract attributions sentences
3
47
93
Michael Jordan
(1) Michael Jordan is a former professional basketball player. (2) He played 15 seasons in the National Basketball Association (NBA) between 1984 and 2003. (3) He won 6 NBA championships with the Chicago Bulls.
Chicago Bulls
(45) The Chicago Bulls are an American professional basketball team based in Chicago. (46) The team was founded on January 16, 1966. (47) The Bulls play their home games at the United Center.
United Center
(93) The United Center is an indoor arena on the Near West Side of Chicago, Illinois. (94) It is home to the Chicago Bulls of the National Basketball Association (NBA) and the Chicago Blackhawks of the National Hockey League (NHL).
Step 3: Prompt LLM for QA pair
Step 4: SynQA training sample
Question: In which arena did Michael Jordan win his NBA championships with the Bulls?
Answer: The United Center
+
Attribution sentences
3
47
93
+
Wikipedia articles
W

Figure 2: **Top:** The SYN-ATT baseline method for synthetic attribution data generation. Given context and question-answer pairs, we prompt an LLM to identify supporting sentences, which are then used to train a smaller attribution model. However, this discriminative approach may yield noisy training data as LLMs are less suited for classification tasks (see §3.3.1). **Bottom:** The SYNQA data generation pipeline leverages LLMs' generative strengths through four steps: (1) collection of Wikipedia articles as source data; (2) extraction of context attributions by creating chains of sentences that form hops between articles; (3) generation of QA pairs by prompting an LLM with only these context attribution sentences; (4) compilation of the final training samples, each containing the generated QA pair, its context attributions, and the original articles enriched with related distractors.

each article in the training sample, we randomly select up to three distractors with the highest semantic similarity to the source articles. These distractors share thematic elements with the source articles, but lack information to answer the questions.

2.2 Advantages of SYNQA

The SYNQA approach has three key advantages: (1) it leverages LLMs' strength in generation rather than classification; (2) creates diverse training samples requiring both dialogue understanding and multi-hop reasoning; and (3) ensures generated questions have clear attribution paths since they are derived from specific context sentences. By generating both entity-centric and dialogue-centric samples, SYNQA produces training data that reflects the variety of real-world QA scenarios, helping models develop robust attribution capabilities, which our experiments demonstrate to generalize across different contexts and domains.

3 Experimental Study

We conduct a comprehensive evaluation across multiple aspects: zero-shot performance, comparison with training on gold attribution data, and generalization to dialogue settings, shedding light on the performance and practical utility of our approach.

3.1 Experimental Setting

We evaluate model performance using precision (P), recall (R), and F1 score. For each sentence in the LLM's output, the context-attribution models identify the set of context sentences that support that output sentence. Precision measures the proportion of predicted attributions that are correct, while recall measures the proportion of ground truth attributions that are successfully identified.

For a fair and comprehensive evaluation, we train all models with a single pass over the training data unless stated otherwise, referring to this setup as **1P** when needed. For a more controlled comparison, some experiments limit the number of training samples each model encounters. Since the synthetic dataset contains approximately 1.0M samples, we allow models to *observe* an equivalent number of samples from the gold training set, ensuring comparable exposure to models trained on data from SYNQA. We refer to this setting as **1M** when necessary. For all models, we fine-tune only the LoRA parameters (alpha=64, rank=32) using a fixed learning rate of 1e-5 and a weight decay of 1e-3.

In-domain datasets: We use SQuAD (Rajpurkar et al., 2016) and HotpotQA (Yang et al., 2018) as our primary in-domain benchmarks.[1] SQuAD provides clear sentence-level evidence for answering questions, serving as a strong baseline for direct attribution. HotpotQA introduces multi-hop reasoning, requiring models to link information across multiple sentences (sometimes from different articles) to identify the correct evidence chain. Additionally, HotpotQA includes distractor documents—closely related yet incorrect sources—posing a more challenging but realistic setting for evaluating attribution performance.

Out-of-domain datasets: To assess generalization beyond the training distribution, we evaluate models on QuAC (Choi et al., 2018), CoQA (Reddy et al., 2018), OR-QuAC (Qu et al., 2020), and DoQA (Campos et al., 2020). These datasets present conversational QA scenarios that differ from SQuAD and HotpotQA. Specifically, QuAC and CoQA introduce multi-turn dialogue structures with coreferences, challenging models to track context across multiple turns. This conversational nature creates a methodological challenge: while these datasets are valuable for evaluating dialogue-based attribution, their reliance on conversation history makes direct comparison with models trained on single-turn QA datasets impossible.

To enable comprehensive evaluation across dialogue QA and single-turn QA, we create two versions of each dataset: (i) a rephrased version using Llama 70B (Dubey et al., 2024) that converts questions into standalone format for fair comparison with models trained on single-turn context attribution (suffixed by "-ST"), and (ii) the original version for assessing dialogue-based attribution.

DoQA extends this challenge further by incorporating domain-specific dialogues (cooking, travel and movies), thus testing the models' adaptability to specialized contexts. OR-QuAC includes context-independent rewrites of the dialogue questions, such that they can be posed in isolation of prior context (i.e., single-turn QA). This enables us to test the models on their capabilities in both single-turn QA and dialogue QA settings.

3.2 Methods

We compare our method (SYNQA) against several baselines, including sentence-encoder-based models, zero-shot instruction-tuned LLMs, and models trained on synthetic and gold context-attribution data. Specifically, we experiment with the following methods:

Sentence-Encoders: We embed each sentence in the context along with the question-answer pair, and select attribution sentences based on cosine similarity with a fixed threshold, tuned on a small validation set.

Zero-shot (L)LMs: We evaluate various instruction-tuned (L)LMs in a zero-shot manner, as such models have been shown to perform well across a range of NLP tasks (Shu et al., 2023; Zhang et al., 2023). During inference, we provide an instruction template describing the task to the LLM (see Appendix C for details).

Ensembles of LLMs: We aggregate the predictions of multiple LLMs through majority voting, selecting attribution sentences that receive consensus from at least 50% of the ensemble. In our experiments, we use Llama8B (Dubey et al., 2024), Mistral7B, and Mistral-Nemo12B (Jiang et al., 2023) as the ensemble constituents.

Models trained on in-domain gold data: Fine-tuning on gold-labeled attribution data provides an upper bound on in-domain performance, helping us assess how well synthetic training data generalizes.

SYN-ATT: SYN-ATT generates synthetic training data by prompting multiple LLMs to perform context attribution in a discriminative manner, aggregating their outputs via majority voting, and training a smaller model on the resulting dataset. To make it a stronger baseline against SYNQA, we give the training data of SQuAD and HotpotQA (the context, questions, and answers) to the LLMs and ask them to perform context attribution (note that we do not use the gold attribution). Finally, we train a model on the generated synthetic data.

SYNQA: We train models using synthetic data generated by our proposed method SYNQA. Note that even though we train models using SYNQA

[1] For some experiments (e.g., in Table 1), these datasets are also *out-of-domain* w.r.t. data generated by SYNQA.

attribution data, we ensure they are not exposed to *any* parts of the evaluation data.[2]

3.3 Results and Discussion

Evaluating our context attribution models requires a multifaceted approach, as performance is influenced by both the quality of training data and the model's ability to generalize beyond in-domain distributions. Therefore, we design our experiments to address five core questions: (i) How well do zero-shot LLMs perform on context-attribution QA tasks (§3.3.1)? (ii) Can models trained on synthetic data generated by SYNQA exceed the performance of models trained on gold context-attribution data (§3.3.2)? (iii) To what extent do models generalize to dialogue settings where in-domain training data is unavailable (§3.3.3)? (iv) How well do models scale in terms of synthetic data quantity generated by SYNQA (§3.3.4)? (v) How do improved context attributions impact the end users' speed and ability to verify questions answering outputs (§3.3.5)?

3.3.1 Comparison to Zero-Shot Models

In Table 1, we present the performance of zero-shot models, and models trained without gold context-attribution data. State-of-the-art sentence-encoder models (e.g., E5) perform relatively poorly, consistent with prior findings (Cohen-Wang et al., 2024). In contrast, LLMs exhibit strong performance, with improvements correlating with model size. Ensembling multiple zero-shot LLMs further enhances performance, leveraging complementary strengths across models, but making the attribution more expensive. We also tested models trained with the discriminative method SYN-ATT. These models significantly outperform their non-fine-tuned counterparts of the same size. However, as postulated, our generative approach SYNQA outperforms SYN-ATT significantly in all but one case. Additionally, SYNQA surpasses zero-shot LLMs that are orders of magnitude larger, showing that we can train a model that is both more accurate and efficient.

3.3.2 Comparison to Models Trained on Gold Attribution Data

In Table 2, we compare models trained on synthetic and gold in-domain context-attribution datasets. As expected, fine-tuning on in-domain gold datasets (SQuAD and HotpotQA) yields highly specialized models that perform well on in-domain data. However, models trained on data obtained by SYNQA exhibit competitive performance on in-domain tasks and consistently surpass in-domain-trained models on out-of-domain datasets. This strong out-of-domain generalization is crucial for practical deployments, where models must handle diverse, previously unseen contexts that often differ substantially from their training data.

3.3.3 Comparison to Zero-Shot and Fine-Tuned Models in Dialogue Contexts

We evaluate dialogue context attribution, for which we do not use any gold in-domain training data (Tab. 3). Here, models must handle follow-up questions that rely on previous turns, often involving coreferences and other dialogue-specific complexities. As expected, zero-shot LLMs exhibit a strong size-performance correlation, with larger models consistently outperforming smaller ones—even those fine-tuned on single-turn question-answer attribution (trained on gold SQuAD and HotpotQA data). However, fine-tuning smaller models with our synthetic data generation strategy leads to superior performance, surpassing both their fine-tuned counterparts and much larger zero-shot LMs. This demonstrates the effectiveness of SYNQA in enhancing context attribution in a dialogue setting and without requiring in-domain supervision.

3.3.4 Scaling Trends and Generalization Performance

Fig. 3a shows F1 scores averaged across datasets, with model size on the x-axis and performance on the y-axis. Models trained on SYNQA-generated data significantly outperform their baseline zero-shot counterparts, while also achieving superior performance compared to zero-shot LLMs that are orders of magnitude larger. This shows our method is highly data-efficient, enabling small models to close the gap with much larger counterparts.

In Figure 3b, we analyze model performance as the quantity of synthetic training data increases, reporting F1 scores separately for in-domain and out-of-domain datasets. As we scale data quantity, performance improves consistently across datasets for isolated context attribution. This trend highlights the scalability of our approach, indicating that further gains can be achieved by increasing synthetic data availability.

[2] We identify data leakage by representing each Wikipedia article as a MinHash signature. Then, for each training Wikipedia article, we retrieve candidates from the testing datasets via Locality Sensitivity Hashing and compute their Jaccard similarity (Dasgupta et al., 2011). We flag as potential leaks pairs exceeding a threshold empirically set to 0.8.

Model	Training data	Squad			Hotpot			Quac-ST			CoQA-ST		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines													
Random	–	19.8	15.4	17.3	4.8	15.2	7.3	5.2	15.1	7.7	7.3	15.1	9.9
E5 ǀ 561M	Zero-shot	38.1	76.5	50.9	12.4	41.4	19.1	65.0	73.8	69.1	61.1	15.2	24.4
HF-SmolLM2 ǀ 365M	Zero-shot	28.1	46.4	35.0	5.1	7.3	6.0	10.6	22.6	14.4	10.6	21.5	14.2
Llama ǀ 1B	Zero-shot	37.5	62.0	46.7	5.3	28.1	8.9	8.8	65.4	15.4	11.9	52.8	19.4
Mistral ǀ 7B	Zero-shot	71.5	94.4	81.4	42.9	42.7	42.8	63.2	88.6	73.8	59.0	72.2	64.9
Llama ǀ 8B	Zero-shot	71.9	96.9	82.6	49.2	52.9	51.0	64.1	92.1	75.6	55.7	76.4	64.4
Mistral-NeMo ǀ 12B	Zero-shot	89.5	94.5	91.8	46.4	47.3	46.8	81.8	85.3	83.5	79.0	67.2	72.6
Ensemble ǀ 27B	Zero-shot	83.1	96.3	89.2	48.1	59.6	53.2	74.8	90.3	81.8	69.5	73.6	71.5
Llama ǀ 70B	Zero-shot	95.3	95.6	95.5	87.6	37.5	52.5	89.7	87.8	88.7	**87.5**	**73.3**	**79.8**
Baselines													
Llama ǀ 1B	Syn-Att (1P)	89.8	96.5	93.0	50.6	58.6	54.3	64.9	91.5	75.9	53.1	75.5	62.3
Llama ǀ 1B	Syn-Att (1M)	84.3	**96.9**	90.2	54.4	58.0	56.1	63.4	92.4	75.2	52.5	77.5	62.6
Ours													
Llama ǀ 1B	SynQA	**96.0**	96.2	**96.1**	**89.6**	**69.4**	**78.2**	**93.3**	**89.1**	**91.1**	<u>82.3</u>	68.5	<u>74.8</u>

Table 1: Comparison of zero-shot models and those trained with synthetic data. Larger zero-shot LMs excel, but our SynQA model outperforms all but one for one dataset while being smaller. **Bold** denotes best method, <u>underline</u> if our method is second best. 1P: models trained with a single pass over the training data. 1M: models trained with 1M samples to match the size of the SynQA data.

Model	Training data	**In-Domain**						**Out-of-Domain**					
		SQuAD			HotpotQA			QuAC-ST			CoQA-ST		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines													
Llama ǀ 1B	Zero-shot	37.5	62.0	46.7	5.3	28.1	8.9	8.8	65.4	15.4	11.9	52.8	19.4
Llama ǀ 1B	SQuAD (1P)	98.4	98.4	98.4	48.7	20.0	28.4	92.6	85.8	89.0	79.9	64.3	71.2
Llama ǀ 1B	HotpotQA (1P)	41.3	87.3	56.0	87.5	79.9	83.5	45.2	89.9	60.1	41.0	70.9	52.0
Llama ǀ 1B	SQuAD & HotpotQA (1P)	98.3	98.3	98.3	**89.7**	78.9	84.0	90.4	90.0	90.2	83.1	68.0	74.8
Llama ǀ 1B	SQuAD & HotpotQA (1M)	**98.3**	**98.4**	**98.3**	87.0	**85.2**	**86.1**	84.0	89.2	86.6	79.2	66.4	72.2
Ours													
Llama ǀ 1B	SynQA	96.0	96.2	96.1	<u>89.6</u>	69.4	78.2	<u>93.3</u>	89.1	<u>91.1</u>	82.3	68.5	<u>74.8</u>
Llama ǀ 1B	SynQA & SQuAD & HotpotQA	<u>98.2</u>	<u>98.3</u>	<u>98.2</u>	89.3	<u>82.4</u>	<u>85.8</u>	**94.5**	**92.7**	**93.6**	**85.5**	**71.0**	**77.6**

Table 2: Comparison of models fine-tuned on synthetic vs. gold in-domain data. Our SynQA approach generalizes better while remaining competitive in-domain. **Bold** denotes best method, <u>underline</u> our method when second best. 1P: models trained with a single pass over the training data. 1M: models trained with 1M samples to match the size of the SynQA data.

3.3.5 User Study: SynQA increases efficiency and accuracy assessment

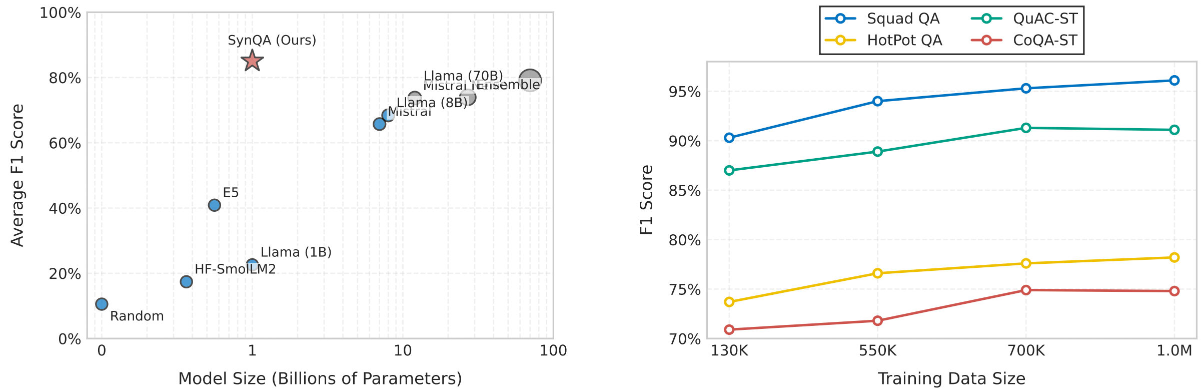
We conducted a user study to evaluate the efficiency and accuracy of verifying the correctness of LLM-generated answers using context attribution. Our hypothesis is that higher-quality context attributions, visualized to guide users, facilitate faster and more accurate verification of LLM outputs. Specifically, in each trial, we presented users with a question, a generated answer, and relevant context, along with attributions visualized as highlights. Their task was to leverage these attributions to judge if the answer was correct w.r.t. a provided context. See Figure 5 in Appendix E.

The study compares three scenarios: (i) **No Alignment:** a baseline condition without context attributions, requiring users to manually read and verify the answer against the entire context; (ii) **Llama 1B (Zero-shot):** context attributions generated by the Llama 1B model were visualized; (iii) **SynQA**: context attributions generated by our approach were visualized.

We employed a within-subjects experimental design for our human evaluation (with 12 participants), ensuring that the same participants evaluate all the aforementioned alignment scenarios, thus requiring fewer participants for reliable results (Greenwald, 1976). However, this can be susceptible to learning effects where participants perform

Model	Training data	Out-of-Domain											
		QuAC			CoQA			OR-QuAC			DoQA		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines													
Llama | 1B	Zero-shot	30.8	45.5	36.8	39.4	37.9	38.6	33.0	46.6	38.6	12.2	22.6	15.9
Mistral | 7B	Zero-shot	76.6	81.8	79.1	67.6	61.3	64.3	82.5	85.1	83.8	74.9	77.9	76.4
Llama | 8B	Zero-shot	84.7	88.8	86.7	79.3	72.0	75.5	88.0	91.3	89.6	77.9	91.4	84.1
Mistral-NeMo | 12B	Zero-shot	85.7	85.4	85.5	81.9	68.4	74.5	88.9	88.8	88.8	86.0	84.2	85.1
Llama | 70B	Zero-shot	88.5	87.7	88.1	**88.3**	**74.9**	**81.1**	81.7	86.3	83.9	85.2	82.0	83.5
Baselines													
Llama | 1B	SQuAD & HotpotQA (1P)	71.3	66.8	69.0	79.0	64.2	70.8	61.6	57.5	59.5	67.4	57.8	62.2
Llama | 1B	SQuAD & HotpotQA (1M)	52.6	49.3	50.9	61.2	50.2	55.2	48.5	44.6	46.5	53.2	49.1	51.1
Ours													
Llama | 1B	SynQA	**91.3**	91.4	91.3	81.7	71.4	76.2	**92.6**	95.3	**94.0**	**86.3**	94.5	**90.2**
Llama | 1B	SynQA & SQuAD & HotpotQA	91.1	**92.2**	**91.7**	82.3	73.2	77.5	90.3	**96.4**	93.2	85.1	**96.0**	**90.2**

Table 3: Context attribution on QuAC, CoQA, OR-Quac, and DoQA (dialogue data); all datasets are out-of-domain. Despite the size advantage of zero-shot LLMs, our SynQA models outperform fine-tuned and larger zero-shot models. **Bold** denotes best method, underline our method when second best. 1P: models trained with a single pass over the training data. 1M: models trained with 1M samples to match the size of the SynQA data.



(a) Model performance vs. size.

(b) F1 score vs. training data size.

Figure 3: Comparison of model performance and scalability. (a) Larger zero-shot models achieve good F1 scores, but our method SynQA (based on Llama 1B) outperforms them while being orders of magnitude smaller. (b) Performance improves consistently with more SynQA training data, highlighting its scalability.

better in later scenarios, because they learned the task from previous examples. To mitigate this, we counterbalanced the scenario order using a Latin Square design (Belz and Kow, 2010; Bradley, 1958), where each alignment scenario appears in each position an equal number of times across all participants. Finally, we randomized the example order within each scenario per participant. For each example, we measured: **verification time** (seconds from display to judgment submission) and **verification accuracy** (binary correct/incorrect judgment).

Results. We observed a clear trend in verification performance across the different attribution settings, with SynQA demonstrating superior effectiveness (Fig. 5). SynQA has the lowest average verification time per example (**148.6** seconds), significantly faster than *No Alignment* (171.8 seconds) and attributions from *Llama 1B* (163.4 seconds). Concurrently, in terms of verification accuracy, SynQA achieved the highest average accuracy (**86.4%**). While *No Alignment* (84.1%) and *Llama 1B (77.3%)* also yielded reasonable accuracy, attributions from SynQA are clearly of higher quality helping users be more accurate.

4 Related Work

We split the related work on context attribution for QA into two categories: (1) in-line citation generation: LLMs are instructed to generate citations along with the generated answer; (2) post-hoc context attribution: perform the attribution *after* the LLM generates the answer. In this section, we outline these works and their differences from our work (for more detailed discussion, see App. A).

4.1 In-line Citation Generation

In this setup, researchers use LLMs to produce in-line citations along with the generated text (Bohnet et al., 2022; Gao et al., 2023b; Huang et al., 2024b). This typically works on paragraph or document level. One line of work focuses on fine-tuning methods for tackling the problem (Gao et al., 2023b; Schimanski et al., 2024; Berchansky et al., 2024; Patel et al., 2024), while another line of work proposes synthetic data generation methods for fine-tuning such models (Huang et al., 2024a,b). Slobodkin et al. (2024) propose a fine-grained task, where the attributions are on sentence level, because such granularity is more useful to human end users. Since generating such in-line citations can result in producing completely made up citations, Yue et al. (2023) propose a task that checks whether in-line generated citations from LLMs are actually attributable or not. Unlike such approaches, we focus on post-hoc context attributions, because this directly predicts a link to a factual source, and therefore avoiding the risk of making up the source.

4.2 Post-hoc Context Attribution

In post-hoc context attribution, the aim is to determine which parts of the context are attributable to an already answered question (Yang et al., 2018). There has been a significant amount of work on training models for the context attribution problem on sentence level for multi-hop QA (Zhang et al., 2024; Ho et al., 2023; Yin et al., 2023; Fu et al., 2021; Tu et al., 2020; Fang et al., 2020). However, they do not investigate this problem in the context of LLMs. Moreover, the methods are constrained *only* to multi-hop QA, and are not tested on broader QA context, such as on dialogue QA. In our work, we propose methods that use LLMs as data generators. This allows us to better generalize and cover multiple QA settings simultaneously, therefore better matching real-world needs.

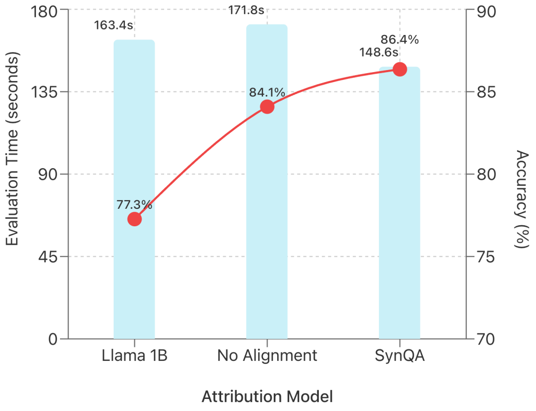


Figure 4: Relationship between Evaluation Time (seconds) and Accuracy (%) for three answer verification settings: *Llama 1B (Zero-shot)*, *No Alignment* and SynQA. SynQA demonstrates the lowest evaluation time and highest accuracy, indicating its superior performance in facilitating efficient and accurate answer verification.

Another line of work focuses on coarse-level granularity and provide attributions either on paragraph level (Rashkin et al., 2023; Menick et al., 2022) or document level (Nakano et al., 2021; Gao et al., 2023a; Buchmann et al., 2024). However, in a user study Slobodkin et al. (2024) observe that such granularity level is not optimal for humans when manually fact-checking LLM-generated content. Their experiments suggest that sentence-level granularity is ideal for humans. This is why we adopt sentence-level granularity in our work, despite this being a harder task. On the other hand, there has been work that focuses on the other extreme: assigning context attributions on sub-sentence level (Cohen-Wang et al., 2024; Phukan et al., 2024). Such methods are computationally expensive and this hinders their practical usability. Our work ensures that models can be run in real-time to make them practical for end users.

5 Conclusion

We investigated the task of context attribution in QA. We focused on approaches that enhance attribution performance without relying on prohibitive human annotations. Our proposed data synthesis strategy, SynQA, enables the generation of high-quality synthetic attribution data, leading to substantial improvements in fine-tuned small models.

Through extensive experiments on six datasets across single turn QA and dialogue QA attribution, we demonstrated that small models fine-tuned with SynQA data (i) significantly outperform models trained on alternative synthetic attributions, (ii) exceed the performance of zero-shot LLMs that are orders of magnitude larger, and (iii) generalize better to out-of-domain distributions compared to models trained on gold in-domain data. These findings suggest that SynQA reduces reliance on large-scale human-labeled datasets, while improving attribution robustness across diverse scenarios.

Finally, our user study validates the practical utility of fine-tuned small models in real-world question-answering applications.

Limitations

While our work demonstrates the effectiveness of SynQA for context attribution in question answering, we leave some important directions for future research. First, all models we train operate exclusively at the sentence level. Even though Slobodkin et al. (2024) found through a user study that sentence-level granularity of context attribution QA is probably the best suited granularity for manual verification of LLM output, this might not always be the optimal granularity for attribution in other tasks or scenarios. Namely, some context elements might be better captured at different levels: e.g., from individual phrases to multi-sentence passages—depending on the semantic structure of the text.

Second, while we evaluated our approach on OR-QuAC, we have not fully explored context attribution in retrieval-augmented generation (RAG) settings with dialogue. This represents a particularly challenging scenario where both the conversational nature of questions and dynamic context updating must be handled simultaneously. Future work should investigate how context attribution models can adapt to streaming contexts when the relevant context continuously evolves throughout a conversation.

Third, we focused primarily on question answering, but context attribution is valuable for many other natural language processing tasks: e.g., in text summarization, attributing summary sentences to source document segments could enhance transparency and fact-checking capabilities. Future research should examine how SynQA's synthetic data generation approach can be adapted for different tasks, potentially revealing task-specific challenges and opportunities for improving attribution mechanisms.

Fourth, our user study (§3.3.5), while providing valuable initial insights into the effectiveness of context attribution to help users verify the LLM model outputs in QA settings, was conducted with a limited sample of 12 participants. A larger-scale study with more participants would strengthen the statistical validity of our findings and potentially reveal more nuanced patterns. Future work should extend this evaluation to a more diverse and larger participant pool, ideally, including users with varying levels of domain expertise and familiarity with language model outputs.

References

R. Anantha, Svitlana Vakulenko, Zhucheng Tu, S. Longpre, Stephen G. Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. In *North American Chapter of the Association for Computational Linguistics*.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, and Peter Izsak. 2024. CoTAR: Chain-of-Thought Attribution Reasoning with Multi-level Granularity. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 236–246.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William Cohen W., Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. *arXiv preprint arXiv:2212.08037*.

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

James V Bradley. 1958. Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. *Journal of the American Statistical Association*, 53(282):525–528.

Samuel Broscheit, Kiril Gashteovski, Yanjie Wang, and Rainer Gemulla. 2020. Can We Predict New Facts with Open Knowledge Graph Embeddings? A Benchmark for Open Link Prediction. In *Association for Computational Linguistics (ACL)*, pages 2296–2308.

Jan Buchmann, Xiao Liu, and Iryna Gurevych. 2024. Attribute or Abstain: Large Language Models as Long Document Assistants. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 8113–8140.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA-Accessing Domain-Specific FAQs via Conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context.

In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. 2024. ContextCite: Attributing Model Generation to Context. *ArXiv*, abs/2409.00729.

Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. ClaimVer: Explainable Claim-Level Verification and Evidence Attribution of Text Through Knowledge Graphs. *ArXiv*, abs/2403.09724.

Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. 2011. Fast Locality-Sensitive Hashing. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1073–1081.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4599–4610.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational agents. In *Conference on Learning Representations (ICLR)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

David Dukić, Kiril Gashteovski, Goran Glavaš, and Jan Snajder. 2024. Leveraging open information extraction for more robust domain transfer of event trigger detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1197–1213.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. Proposition-level Clustering for Multi-document Summarization. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1765–1779.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source Proposition-level Alignment: Task, Datasets and Supervised Baseline. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 310–322.

Ori Ernst, Ori Shapira, Aviv Slobodkin, Sharon Adar, Mohit Bansal, Jacob Goldberger, Ran Levy, and Ido Dagan. 2024. The Power of Summary-Source Alignments. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 6527–6548.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical Graph Network for Multi-hop Question Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing Complex Questions Makes Multi-Hop QA Easier and more Interpretable. In *Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)*, pages 169–180.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, N. Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling Large Language Models to Generate Text with Citations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6465–6488.

Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling, and Christian Meilicke. 2020. On Aligning OpenIE Extractions with Knowledge Bases: A Case Study. In *Proceedings of the Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP@ACL*, pages 143–154.

Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. 2019. OPIEC: An Open Information Extraction Corpus. In *Conference on Automated Knowledge Base Construction (AKBC)*.

Anthony G Greenwald. 1976. Within-Subjects Designs: To Use or not to Use? *Psychological Bulletin*, 83(2):314.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2023. Analyzing the Effectiveness of the Underlying Reasoning Tasks in Multi-hop Question Answering. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL Findings)*, pages 1133–1150.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024a. Learning Fine-Grained Grounded Citations for Attributed Large Language Models. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 14095–14113.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024b. Advancing Large Language Model Attribution through Self-Improving. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, Giuseppe Marra Pascal Hitzler, Pasquale Minervini, Martin Mundt, Axel-Cyrille Ngonga Ngomo, Alessandro Oltramari, Gabriella Pasi, Zeynep G. Saribatur, Luciano Serafini, John Shawe-Taylor, Vered Shwartz, Gabriella Skitalinskaya, Clemens Stachl, Gido M. van de Ven, and Thomas Villmann. 2024. Aligning Generalisation Between Humans and Machines. *arXiv preprint arXiv:2411.15626*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *ArXiv*, abs/2310.06825.

Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022a. Human-Centric Research for NLP: Towards a Definition and Guiding Questions. *arXiv preprint arXiv:2207.04447*.

Bhushan Kotnis, Kiril Gashteovski, and Carolin Lawrence. 2023. Open Information Extraction from Low Resource Languages. US Patent 11,741,318.

Bhushan Kotnis, Kiril Gashteovski, Daniel Rubio, Ammar Shaker, Vanesa Rodriguez-Tembras, Makoto Takamoto, Mathias Niepert, and Carolin Lawrence. 2022b. MILIE: Modular & Iterative Multilingual Open Information Extraction. In *Association for Computational Linguistics (ACL)*, pages 6939–6950.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1642–1661.

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*, abs/2311.03731.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How Hard is Automatic Attribution Evaluation? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14919–14935.

Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella, and Anandhavelu Natarajan. 2024. Presentations are not always Linear! GNN Meets LLM for Document-to-Presentation Transformation with Attribution. In *Conference on Empirical Methods in Natural Language Processing (EMNLP, Findings)*, pages 15948–15962.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching Language Models to Support Answers with Verified Quotes. *ArXiv*, abs/2203.11147.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted Question-answering with Human Feedback. *ArXiv*, abs/2112.09332.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290.

Federico Nanni, Jingyi Zhang, Ferdinand Betz, and Kiril Gashteovski. 2019. EAL: A Toolkit and Dataset for Entity-Aspect Linking. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 430–431. IEEE.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 9844–9855.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards improved Multi-source Attribution for Long-form Answer Generation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3906–3919.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. Peering into the Mind of Language Models: An Approach for Attribution in Contextual Question Answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11481–11495.

Jirui Qi, Gabriele Sarti, Raquel Fern'andez, and Arianna Bisazza. 2024. Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 6037–6053.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 539–548.

Gorjan Radevski, Kiril Gashteovski, Chia-Chien Hung, Carolin Lawrence, and Goran Glavaš. 2023. Linking Surface Facts to Large-Scale Knowledge Graphs. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7189–7207.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Vasan Srinivavsan. 2024. Enhancing Post-Hoc Attributions in Long Document Comprehension via Coarse Grained Answer Decomposition. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 17790–17806.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics (TACL)*, 7:249–266.

Wiem Ben Rim, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Carolin Lawrence, Jürgen Quittek, and Sascha Saralajew. 2024. A Human-Centric Assessment of the Usefulness of Attribution Methods in Computer Vision. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 20–37. Springer.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards Faithful and Robust LLM Specialists for Evidence-Based Question-Answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1913–1931.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the Exploitability of Instruction Tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:61836–61856.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute First, then Generate: Locally-attributable Grounded Text Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3309–3344.

Johanne R Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users really Ask Large Language Models? An initial Log Analysis of Google BARD Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2703–2707.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, Answer and Explain: Interpretable Multi-Hop Reading Comprehension over Multiple Documents. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 9073–9080.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering. *Transactions of the Association for Computational Linguistics (TACL)*, 12:321–333.

Denny Vrandečić and Markus Krötzsch. 2014. WikiData: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57(10):78–85.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533*.

Zhao Xu, Wiem Ben Rim, Kiril Gashteovski, Timo Sztyler, and Carolin Lawrence. 2024. A Human-Centric Evaluation Platform for Explainable Knowledge Graph Completion. In *European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pages 18–26.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2013–2018.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhangyue Yin, Yuxin Wang, Xiannian Hu, Yiguang Wu, Hang Yan, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2023. Rethinking Label Smoothing on Multi-hop Question Answering. In *China National Conference on Chinese Computational Linguistics*, pages 72–87. Springer.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In *Findings of ACL: Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)*, pages 4615–4635.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. 2024. End-to-End Beam Retrieval for Multi-Hop Question Answering. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1718–1731.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction Tuning for Large Language Models: A survey. *arXiv preprint arXiv:2308.10792*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Conference on Learning Representations (ICLR)*.

A Comprehensive Discussion on Related Work

We split the related work papers into several categories: tasks, datasets, methods and metrics.

A.1 Context Attribution Tasks

Attributable to Identified Sources (AIS): given a generative text t_g and a context text t_c, is t_g attributable to t_c? Rashkin et al. (2023) propose a manual framework that defines the AIS task and evaluates the AIS scores across several NLP tasks, namely conversational QA (Anantha et al., 2020; Dinan et al., 2019), text summarization (Nallapati et al., 2016) and table-to-text (Parikh et al., 2020). Our work differentiates in three important aspects: (1) we focus on a broader QA setup (i.e., single-question QA and conversational QA), which makes our work a subset of the broader AIS task; (2) we focus on more fine-grained level: our attributions are not on the entire text level, but rather on a sentence level, which has been shown in user studies to be more useful to end-users (Slobodkin et al., 2024); (3) the AIS task entails a manual evaluation framework, while our work provides automatic evaluation with golden data.

In-line Citation Generation: uses LLMs to produce in-line citations along with the generated text (Bohnet et al., 2022; Gao et al., 2023b; Huang et al., 2024b). This typically works on paragraph or document level. Slobodkin et al. (2024) propose a fine-grained task, where the attributions are on sentence level, because such granularity is more useful to human end users. Because generating such in-line citations can result in producing completely made up citations, Yue et al. (2023) propose a task that checks whether the in-line generated citations from LLMs are actually attributable or not. Instead of using binary attributable/non-attributable labels (like with AIS), they propose more fine-grained labels for this problem: attributable, extrapolatory, contradictory and non-attributable. Contrary to such approaches, our work focuses on post-hoc context attributions: given an answer to a question, find the sentences in the context that support the factuality of the answer.

Post-hoc Attribution: determines which parts of the context are attributable to an already answered question (Yang et al., 2018). Within the post-hoc context, there are two other subcategorizations of the task: *contributive* and *corroborative post-hoc attribution* (Cohen-Wang et al., 2024).

Post-hoc Attribution (Contributive): ContextCite (Cohen-Wang et al., 2024) and Mirage (Qi et al., 2024) define a post-hoc task that aims at discovering which parts of the context *caused* the LLM to generate the particular response. Their evaluation methods, however, are based on proxy metrics that do not rely on golden annotations, while in our work we rely on automatic annotations that rely on golden data.

Post-hoc Attribution (Corroborative): this task is similar to contributive post-hoc attribution. The difference is that the constraint for causality is not necessarily enforced, but should support the factuality of the statement (Cohen-Wang et al., 2024). Many works are based on coarse-grained level and provide attributions on either paragraph level (Menick et al., 2022), document level (Nakano et al., 2021) or on multi-document level, where they have a RAG component that retrieves the documents that are potentially attributable (Gao et al., 2023a; Buchmann et al., 2024).

Context Attributions to other Modalities. Other line of work maps the attributions to other modalities, such as knowledge graphs Dammu et al. (2024). Similarly, Maheshwari et al. (2024) take multi document collection as input, construct a graph of narratives, and then generate a presentation (i.e., slides) for the topic, along with attributions from the generated content of the slides with the original documents. We do not investigate such cases, and focus on attributing answers to sentences within the user-provided context.

Post-hoc Attribution for Text Summarization. Ernst et al. (2021) proposed a task, dataset and baseline model (dubbed SuperPAL) for detecting attributions for text summarization. In a followup work, Ernst et al. (2022) an extension of the task, this time for clustering propositions for text summarization and Ernst et al. (2024) extend this to multi-document summarization. Krishna et al. (2023) investigate whether such text summarization alignments are helpful for humans. In our work, we focus on the question answering (QA) task.

A.2 Datasets

AIS. Rashkin et al. (2023) proposed the AIS dataset, which contains three tasks: question answering, table-to-text and text summarization. Here, for each data point, there is a query and an

LLM-generated response, along with label by humans whether it is fully attributable or not. This data is on paragraph and document level, and lacks the granularity of a sentence level. Therefore, we do not use it in our work.

HotpotQA. With HotpotQA (Yang et al., 2018), the authors propose an explainable multi-hop QA dataset. The dataset also contains attribution links (i.e., explanations) for the answers: spans of text that belong to the input context, which are supporting the statement in the answer. The authors set up baselines for measuring the ability of attributions of models on sentence level, which is in line with what we do. In our work, we integrated HotpotQA as part of our setup for both training and testing.

AttributionBench. This is a benchmark for attribution evaluation of LLM generated content (Li et al., 2024). In particular, the benchmark assesses whether the assigned attribution on a generated text is actually attributable. In particular, given a query, response set $\mathcal{R}$ (containing claims) and evidence set E, the task is to label as "attributable" or "not attributable" every claim against E. This work operates on a coarse-grained level (paragraphs or whole documents). Similarly, Yue et al. (2023) proposed another dataset for evaluating attribution of LLM-generated text, same on paragraph level. In our work, we focus on sentence-level context attribution. Therefore, we did not include this dataset in our work.

Conversational QA. CoQA (Reddy et al., 2018) is a conversational QA dataset, which contains a context, questions-answer pairs between two people (teacher and student) and sentence-level supporting evidence for the context. We use this dataset in our evaluation to test the out-of-domain capabilities of LLMs for context attribution. We also use QuAC (Choi et al., 2018) and ORConvQA (Qu et al., 2020), which are conversational QA datasets similar to CoQA.

QASPER. This dataset is from the scientific domain (Dasigi et al., 2021). The dataset contains title and an abstract of a paper, question and answer about the content. This data has information on paragraph level, not on sentence level. Therefore, we do not use it in our experiments.

WikiQA. Dammu et al. (2024) use WikiQA (Yang et al., 2015), because it's Wikipedia-based dataset, which can be linked to Wikipedia-derived KG like Wikidata (Vrandečić and Krötzsch, 2014). In our work, we focus only on text modality, which is why we do not include this dataset into our evaluation.

WikiNLP. The WikiNLP dataset (Gashteovski et al., 2019) is a dataset that contains the entire English Wikipedia, along with linguistic annotations (e.g., POS tags, dependency parse trees, etc.) and semantic annotations (e.g., NER tags and entity links). We use this dataset for the synthetic data generation, because it keeps the linked information from the Wikipedia articles, which are annotated by humans. For this reason, the dataset has been used in wide range of tasks in research, mostly for information extraction (Dukić et al., 2024; Kotnis et al., 2023, 2022b; Gashteovski et al., 2020), but also for other tasks such as clustering (Viswanathan et al., 2024), open link prediction (Broscheit et al., 2020) and entity linking (Nanni et al., 2019; Radevski et al., 2023)

A.3 Metrics and Evaluation

AIS. The AIS framework (Rashkin et al., 2023) is human annotation framework. Given a generated text chunk and a context chunk (this can be sentence, paragraph or document), a human evaluated whether the generated text chunk is fully attributible or not. It is basically a binary classification problem. In their data, the authors focus on document level granularity, which is not useful for humans. In our setup, we check for each sentence in the context if it supports the answer.

AutoAIS. To evaluate the attributed information, Slobodkin et al. (2024) use **AutoAIS metric**: an NLI-based scoring. Prior studies have shown that this metric highly correlated with human annotations (Bohnet et al., 2022; Gao et al., 2023b). This is an extension to the AIS metric. We do not use proxy metrics, but rely on golden annotations by humans.

AttrScore. With AttrScore, Yue et al. (2023) consider LLM generated content on the one hand and citation documents on the other hand. Then, the score evaluates whether a provided citation is attributable, extrapolatory, contradictory or non-attributable. Essentially, it is an extension of AIS, such that it provides more fine grained labels for the provided citations. The AttrScore is basically a fine-tuned LLM that provides these scores.

Unsupervised Metrics. ContextCite proposed the Top-k-drop and LDS metric to evaluate the causal post-hoc attribution. These metrics do not require labeled data. Berchansky et al. (2024) uses ROUGE and BERTScore to evaluate their results. We do not use unsupervised metrics and rely on automated evaluation with golden annotations by humans.

A.4 Methods

Multihop QA. There has been significant amount of work on tackling the context attribution problem on sentence level for multi-hop QA (Zhang et al., 2024; Ho et al., 2023; Yin et al., 2023; Fu et al., 2021; Tu et al., 2020; Fang et al., 2020). While we also investigate this problem, in contrast to our work, these works focus *only* on the multihop QA task. In our work, we also explore other QA setups, including conversational QA with different domains. Moreover, these papers do not investigate the capabilities of LLMs about the context attribution problem, but rather are proposing specific methods that are tailor made for the multi-hop QA problem, which involves both answering the questions and providing supporting sentences to the answers.

In-line Citation Generation. Another line of work focuses on guiding LLMs to generate in-line citations along with the generated text (Li et al., 2023). Slobodkin et al. (2024) tackle this problem on a sentence level, but do not investigate the post-hoc context attribution case. Moreover, they rely on proxy metrics such as AutoAIS (Gao et al., 2023a) and BERTScore (Zhang et al., 2020). Similarly, Bohnet et al. (2022) proposes methods for in-line citation generation, but this work is more coarce grained and focuses on paragraph and document level. They also report their findings on proxy metrics. Their method is based on retrieval and they do not investigate the LLMs capabilities thoroughly. Gao et al. (2023b) assign citations to LLM generated content, where they retrieve the information from a large collection of documents (also, it's on paragraph and document level, not on sentence level). START (Huang et al., 2024b) propose a data synthetic generation method for in-line citation generation on document level, where each citation refers to an entire document. FRONT (Huang et al., 2024a) also investigates synthetic data generation of in-line citation generation, where the citations assigned to the sentences in the output are entire documents. Similarly, Schimanski et al. (2024) propose a synthetic data generation pipeline for fine tuning models that solve the same problem. Berchansky et al. (2024) use Chain-of-Thought approaches and fine-tuning smaller LLMs in order to solve this problem. Patel et al. (2024) also fine-tune a model specifically for this task, and the attributions are on paragraph level.

Post-hoc Context Attribution. Ramu et al. (2024) propose template-based in-context learning method for post-hoc context attribution. In particular, they use standard retrievers as a first step to pre-rank the text (e.g., BM25 and dual encoders (Ni et al., 2022)) and then they use LLMs to calssify (i.e., rerank) the relevant sentences. ContextCite (Cohen-Wang et al., 2024) uses ablation-based methods to infer the attributions of post-hoc generated text.

User Study. Recent work has called for a more human-centric research in NLP (Kotnis et al., 2022a). In such work, the idea is to involve the user (i.e., the final stakeholder) in the process of research, which is typically done with some forms of user studies (Rim et al., 2024; Xu et al., 2024; Ilievski et al., 2024). In this spirit, we want To verify whether our models are useful for end users. To this end, we performed a user study, whereas users were asked to solve a fact checking task with the use of our context attribution model. We found that our approach does indeed make human end-users faster in performing the manual fact-checking task (for details of our user study, see Section 3.3.5).

B Method for Synthetic Data Generation

B.1 Multi-hop Generation of Attribution Data

To generate synthetic data with the use of Wikipedia, we use the WikiNLP dataset (Gashteovski et al., 2019). It contains the text from all Wikipedia articles along with annotations for links within the text that link to other Wikipedia articles. The main idea is to use the links in order to imitate reasoning hops across different (related) articles. Therefore, we filter out all articles that either do not contain links or that contain links to articles that do not contain links. Finally, for each article, we use only the first paragraph, because this is considered to be the paragraph that contains the most "definitional information" (Bovi et al., 2015); i.e., information that precisely describes the target concept of the article and contains the most important information about it.

Then, for each article, we randomly select a sentence that contains at least one link to another Wikipedia article.[3] Each of the sentences that we sample serve as ground truths for the context attribution. With these sentences, we then prompt an LLM to generate a question-answer pair.

B.2 Question-Answer Pairs Generation

Using the multi-hop chain of sentences, we prompt an LLM (see §C) by providing it *only* the formed chain as evidence. The LLM generates a question-answer pair that must be answered using the information in these supporting sentences, ensuring the pairs are grounded in the provided evidence.

B.3 Distractors Mining

In realistic scenarios, whether the context is user-provided or retrieved through RAG, the system typically encounters multiple context documents that are highly similar to those containing the evidence sentences. To bridge this gap between our synthetically generated training data using SynQA and the data models encounter "in the wild", we augment each training sample with hard negative distractor articles. We obtain embeddings using E5 (Wang et al., 2022) for each Wikipedia article in our collection. Then, for each article containing a supporting sentence for the question-answer pair, we randomly sample up to three distractor articles that share semantic similarity with the ground truth article. This process increases the difficulty of the training data, producing models better equipped to handle diverse testing scenarios.

B.4 Comparison to HotpotQA

Although our method is inspired by HotpotQA (Yang et al., 2018), note that we do not aim to recreate the HotpotQA dataset. Our method has significant differences, which result in both much higher amount of data and in higher domain variability.

Particularly, their method is more curated and involves humans in multiple steps. First, the authors manually select the target entities (and, with that, the target articles from which the annotators create the question and answer pairs). The reason for this is because many highly specialized articles—e.g., the article for IPv4 protocol—will not be suitable for crowd-workers to both identify meaningful questions and provide answers for those questions. Our approach does not have this constraint and, therefore, produces data that has much higher domain variability.

Second, their method uses mechanical Turk workers to annotate the questions, answers and attribution sentences. In our case, we automatically select the hopped sentences (which serve as gold context attribution data), and then we use these sentences to generate question-answer pairs with an LLM.

Third, while HotpotQA always enforces multi-hop QA pairs, we do not instruct the LLM to do that. Rather, we first allow the LLM to decide whether generating such multihop QA pair is actually possible for the incoming context attribution sentences. If so, then the LLM generates multihop QA pairs. Otherwise, it generates direct QA pairs that do not need hops; i.e., QA pairs like in SQuAD (Rajpurkar et al., 2016).

Fourth, the HotpotQA annotation method does not allow for dialogue QA. In our method, we also create dialogue multi-hop data.

With these differences in mind, we showed that, compared to HotpotQA, our data generation method exhibits the following advantages: (1) our method generates data with higher domain variability; (2) our method goes beyond multi-hop QA and also generates direct QA pairs (like SQuAD) as well as dialogue QA data; (3) we generate the data in completely automatic manner without the involvement of humans.

C Prompts to generate SynQA synthetic training data

In order to generate the question-answer pairs, we provide Llama 70B with the following prompts:

SYSTEM PROMPT
You are tasked with generating a concise and focused question-answer pair using information from provided Wikipedia sentences. Follow these instructions carefully:
1. You will be provided with multiple Wikipedia articles, each containing:
- The title of the article.
- One specific sentence from the article.
2. Your goal is to generate a **short, factual question** and a **concise answer**, ensuring:

[3] To make sure we have multi-hop scenario, we also check if the other Wikipedia article also contains at least one valid link to another Wikipedia article.

- The question-answer pair is grounded in the provided sentences.
- The reasoning is logical, clear, and references all sentences used.

3. **Key Constraints**:
- Questions must address a **single coherent topic** or concept that can be logically inferred from the provided sentences.
- Avoid combining unrelated pieces of information into a single question.
- The `"reasoning"` must explain how each sentence in the `"ids"` field contributes to answering the question but should remain **brief** and **to the point**.

4. Aim for **brevity**:
- Questions should be concise and avoid unnecessary details.
- Answers should be short, typically no more than one sentence.
- Keep the reasoning concise, focusing only on the necessary logical connections.

5. Multi-hop reasoning is encouraged but must be natural and focused:
- Combine information only when it is logical and directly relevant to the question.
- Do not create overly complex questions that combine weakly related information.

6. Provide your response in **raw JSON format** with the following keys:
- `"question"`: A concise and clear question string.
- `"answer"`: A short and factual answer string.
- `"ids"`: A list of JSON-compatible arrays (e.g., `[[0, 0], [1, 0]]`) representing the indices of all sentences used to generate the question-answer pair.
- `"reasoning"`: A brief explanation of how **each sentence in `"ids"`** was used to generate the question-answer pair.

Important Notes:
- Ensure the question-answer pair is entirely self-contained and logically consistent.
- Do not include unnecessary or weakly related information in the question or answer.
- Avoid introducing information not present in the provided sentences.
- Do not include additional formatting, explanations, or markdown in your response.

USER PROMPT
Here are the titles and sentences:
Title: [First Article Title]
[0, 0] [First sentence from the article]
Title: [Second Article Title]
[1, 0] [Second sentence from the article]
Title: [Third Article Title]
[2, 0] [Third sentence from the article]
Use the provided sentences to generate a question-answer pair following the specified guidelines. Respond **only in raw JSON** with no additional formatting or markdown.

Given the **SYSTEM** and the **USER** prompt, the LLM is generating the question-answer pair, which when combined with the full articles, yields a single SYNQA training data sample.

Should we want to generate a SYNQA dialog training data sample, we make the prompts a bit simpler:

SYSTEM PROMPT
You are an AI assistant that generates structured question-answer pairs based on a passage. Your goal is to create meaningful, factual, and reasoning-based questions that require connecting multiple sentences.
Follow these strict guidelines:
- Format the output as a **valid JSON array**, where each item has:
- `"question"`: A clear, concise question.
- `"answer"`: A short, factual response.
- `"sentence_numbers"`: A list of integers pointing to **all** relevant supporting sentences.
- **Ensure questions are generated in a random sentence order** (not sequential).
- Some questions **must reference multiple sentences** for reasoning.
- Some sentences should be **reused** across multiple questions.
- **Later questions should rely on earlier information** and use pronouns or indirect references to maintain logical flow.
- Introduce a mix of **fact-based, causal, and inference questions**.
- Avoid introducing **information not present in the passage**.
- Ensure **all relevant sentences are cited** for each answer.

> Your response must be **valid JSON** containing 5 to 10 question-answer pairs.

and the user prompt:

> **USER PROMPT**
> Here is a passage:
> Title: [Title of the passage]
> 0. [First sentence of the passage]
> 1. [Second sentence of the passage]
> 2. [Third sentence of the passage]
> 3. [Fourth sentence of the passage]
> ...
> Generate structured question-answer pairs following these constraints:
> - **Return output in JSON format only**: `["question": "...", "answer": "...", "sentence_numbers": [..], ...]`
> - Use **random sentence order**, not sequential.
> - Some questions should require **multiple sentences**.
> - Some sentences should be **reused** across different Q&A pairs.
> - **Later questions must reference earlier ones** using pronouns or indirect mentions.
> - **Include a mix of question types**:
> - Factual questions that can be answered directly from the passage.
> - Causal questions that require understanding relationships between sentences.
> - Inference-based questions that require implicit reasoning.
> - Ensure **sentence numbers fully cover the reasoning required**.
> Return **only** JSON, with no extra text.

D Zero-shot models

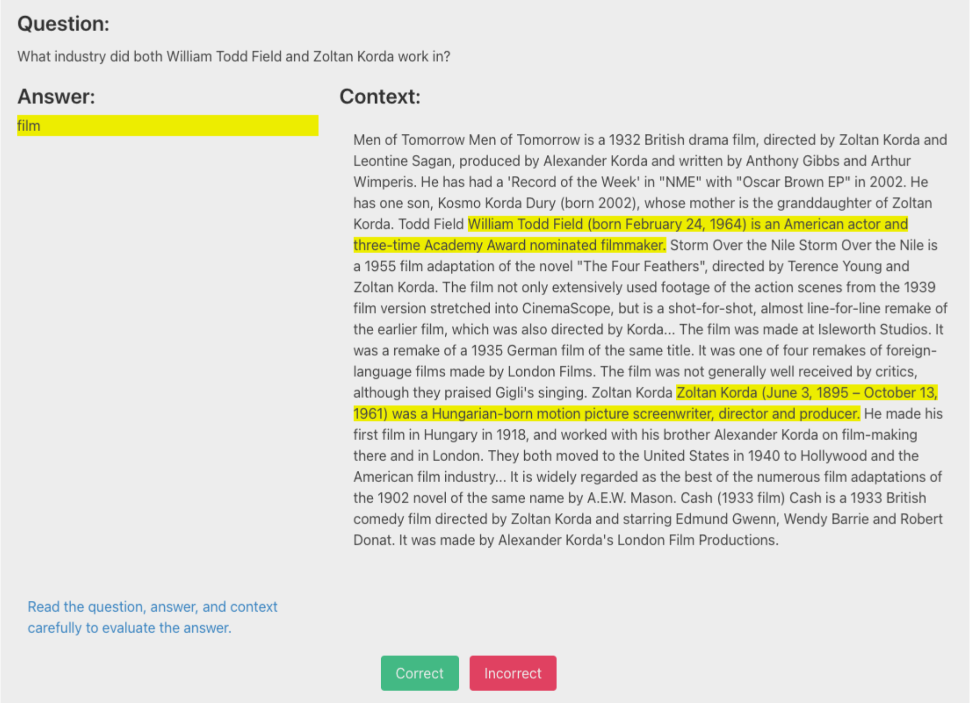
In order to obtain context attributions with the instruction tuned LLM (Jiang et al., 2023; Dubey et al., 2024), we use the following prompts:

> **SYSTEM PROMPT**
> You are an AI assistant that identifies the sentence(s) in a provided context document most relevant for answering a specific question. Your task is to select only the sentence(s) containing the explicit information needed to answer the question accurately, without adding extra context.

> **USER PROMPT**
> Context Document:
> [numbered sentences from the context]
> Question: [query text]
> Answer: [answer text]
> Based on the context document, identify the sentence number(s) from the following choices: [list of numbers]. Select only the sentence(s) that contain explicit information needed to answer the question directly.
> Answer only with the corresponding number(s) in parentheses, without additional explanation.

E User Study

An example of the attribution scenario evaluated in our user study. See Figure 5 for details.



Question:

What industry did both William Todd Field and Zoltan Korda work in?

Answer:

film

Context:

Men of Tomorrow Men of Tomorrow is a 1932 British drama film, directed by Zoltan Korda and Leontine Sagan, produced by Alexander Korda and written by Anthony Gibbs and Arthur Wimperis. He has had a 'Record of the Week' in "NME" with "Oscar Brown EP" in 2002. He has one son, Kosmo Korda Dury (born 2002), whose mother is the granddaughter of Zoltan Korda. Todd Field William Todd Field (born February 24, 1964) is an American actor and three-time Academy Award nominated filmmaker. Storm Over the Nile Storm Over the Nile is a 1955 film adaptation of the novel "The Four Feathers", directed by Terence Young and Zoltan Korda. The film not only extensively used footage of the action scenes from the 1939 film version stretched into CinemaScope, but is a shot-for-shot, almost line-for-line remake of the earlier film, which was also directed by Korda... The film was made at Isleworth Studios. It was a remake of a 1935 German film of the same title. It was one of four remakes of foreign-language films made by London Films. The film was not generally well received by critics, although they praised Gigli's singing. Zoltan Korda Zoltan Korda (June 3, 1895 – October 13, 1961) was a Hungarian-born motion picture screenwriter, director and producer. He made his first film in Hungary in 1918, and worked with his brother Alexander Korda on film-making there and in London. They both moved to the United States in 1940 to Hollywood and the American film industry... It is widely regarded as the best of the numerous film adaptations of the 1902 novel of the same name by A.E.W. Mason. Cash (1933 film) Cash is a 1933 British comedy film directed by Zoltan Korda and starring Edmund Gwenn, Wendy Barrie and Robert Donat. It was made by Alexander Korda's London Film Productions.

Read the question, answer, and context carefully to evaluate the answer.

Figure 5: An example of the attribution scenario evaluated in our user study. Both the answer and the context attributions are highlighted to help the user verify the correctness of the answer. In the absence of highlights, the user is instructed to read the entire context. This example showcases a practical application of context attribution in real world interactions with LLM generated content.