Effective Large Language Model Adaptation for Improved Grounding and Citation Generation

Xi Ye[♢*] **Ruoxi Sun**[♠] **Sercan Ö. Arık**[♠] **Tomas Pfister**[♠]
[♢] The University of Texas at Austin [♠] Google Cloud AI
[♢]xiye@cs.utexas.edu
[♠]{ruoxis,soarik,tpfister}@google.com

Abstract

Large language models (LLMs) have achieved remarkable advancements in natural language understanding and generation. However, one major issue towards their widespread deployment in the real world is that they can generate "hallucinated" answers that are not factual. Towards this end, this paper focuses on improving LLMs by grounding their responses in retrieved passages and by providing citations. We propose a new framework, *AGREE*, **A**daptation for **GR**ounding **E**nhanc**E**ment, that improves the grounding from a holistic perspective. Our framework tunes LLMs to self-ground the claims in their responses and provide accurate citations to retrieved documents. This tuning on top of the pre-trained LLMs requires well-grounded responses (with citations) for paired queries, for which we introduce a method that can automatically construct such data from unlabeled queries. The self-grounding capability of tuned LLMs further grants them a test-time adaptation (TTA) capability that can actively retrieve passages to support the claims that have not been grounded, which iteratively improves the responses of LLMs. Across five datasets and two LLMs, our results show that the proposed tuning-based AGREE framework generates superior grounded responses with more accurate citations compared to prompting-based approaches and post-hoc citing-based approaches.

1 Introduction

Recent advancements in large language models (LLMs) have yielded demonstrably groundbreaking capabilities in natural language processing (NLP) (Brown et al., 2020; Chowdhery et al., 2022). Their ability to understand, generate, and manipulate text at unprecedented scales and depths has established them as a transformative force within the burgeoning field of artificial intelligence, poised to significantly impact our increasingly data-driven world. Despite their widely spread adoption, one prominent issue of LLMs is that in certain scenarios they hallucinate: they generate plausible-sounding but nonfactual information (Maynez et al., 2020; Ji et al., 2023; Menick et al., 2022), limiting their the applicability in real-world settings. To mitigate hallucinations, solutions generally rely on grounding the claims in LLM-generated responses to supported passages by providing an attribution report (Rashkin et al., 2023; Bohnet et al., 2022; Gao et al., 2023a) or adding citations to the claims (Liu et al., 2023; Gao et al., 2023b; Huang and Chang, 2023).

There has been a growing amount of interest in making LLM-generated responses more trustworthy by grounding and adding citations. One line of work uses instruction tuning (Kamalloo et al., 2023) or in-context learning (Gao et al., 2023b) to instruct LLMs to generate grounded responses with citations to retrieved passages, following the retrieval-augmented generation (Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020) framework. As LLMs are required to perform this challenging task from just instructions and few-shot demonstrations, such directions often lead to mediocre grounding quality (Gao et al., 2023b). Another line of work is on post-hoc citing (Gao et al., 2023a; Chen et al., 2023), which links support passages to the claims in responses using a natural language inference (NLI) model. This paradigm heavily relies on LLMs' parametric knowledge and might not extend well to less-known knowledge (Sun et al., 2023).

We propose a new learning-based framework, **AGREE**, **A**daptation of LLMs for **GR**ounding **E**nhanc**E**ment. As shown in Fig. 1, our framework fine-tunes LLMs to generate citations, as opposed to prompting or relying on an external NLI model used in a post-hoc way. At the training

[*] Work done during an internship at Google Cloud AI.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6237–6251

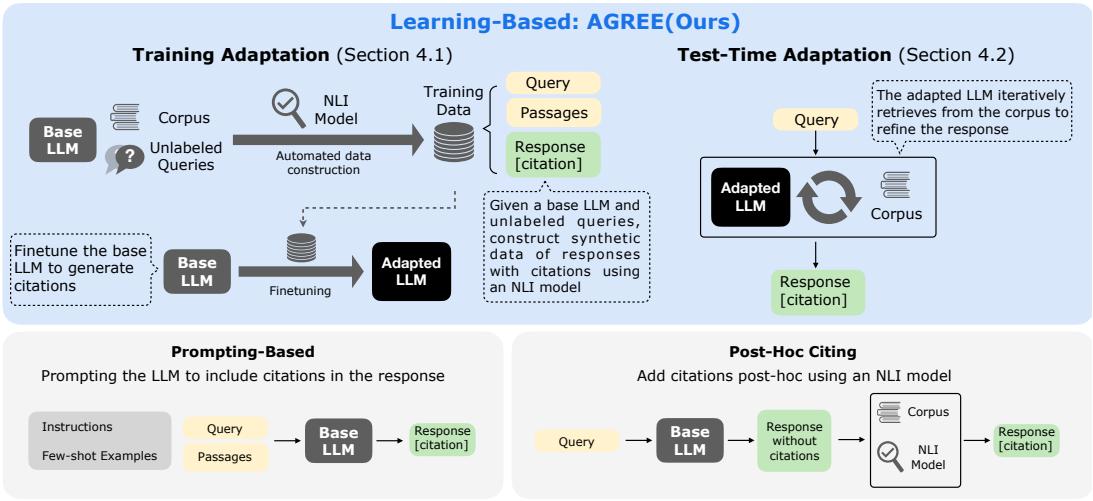

Figure 1: Our framework, AGREE, combines tuning (Section 4.1) and test time adaptation (Section 4.2) for better attribution and citation generation.

phase, AGREE collects well-grounded responses for unlabelled queries automatically from a base LLM with the help of an NLI model. Next, the collected data are used for supervising LLMs to generate grounded responses based on the retrieved passages as well as include citations in their responses. As a test-time approach, we propose an iterative inference strategy that allows LLMs to seek for additional information based on the self-grounding evaluation so as to refine its response. The tuning and test-time adaptation together enable LLMs to effectively and efficiently ground their responses in the corpus. We apply AGREE framework to adapt an API-based LLM, `text-bison`, and an open LLM, `llama-2-13b`, with training data collected using unlabelled queries from three datasets. We conduct evaluation on both in-domain and out-of-distribution datasets, comparing the proposed AGREE framework against competitive in-context learning and post-hoc citing baselines. The experimental results highlight that AGREE framework successfully improves grounding, in citation recall & precision, compared to the baselines by a substantial margin (generally more than 20%). We find LLMs can learn to add accurate citations to their responses with our carefully designed tuning mechanisms. Furthermore, the improvements in grounding quality achieved by tuning using certain datasets can generalize well across domains. To summarize, our main contributions include:

- A learning-based approach that adapts a base LLM to include accurate citations in its response, leveraging automatically created data;
- A test-time adaptation (TTA) method that iteratively improves responses of LLMs based on the citation information;
- Extensive experiments on two LLMs over five datasets demonstrating the effectiveness of the proposed AGREE framework for improving grounding and citation generation.

2 Related Work

Hallucination is a prevalent issue for generative language models on many tasks (Maynez et al., 2020; Raunak et al., 2021; Dziri et al., 2021; Ji et al., 2023; Ye and Durrett, 2022; Tang et al., 2023; Huang and Chang, 2023). It has been evaluated in different ways, investigating the grounding in generated responses (Bohnet et al., 2022; Rashkin et al., 2023; Min et al., 2023; Yue et al., 2023).

Various approaches have been proposed to mitigate hallucination and improve the factuality of LLM-generated responses. Among these, our work particularly focuses on providing citations to attributable information source (Liu et al., 2023; Gao et al., 2023b). Unlike existing work that largely relies on zero-shot prompting or few-shot prompting (Kamalloo et al., 2023; Gao et al., 2023b) or use an additional NLI model (Gao et al., 2023a; Chen et al., 2023) to add citations, we propose a learning-based approach that tunes LLMs to generate better-grounded responses supported with citations.

More broadly, recent work also investigates methods for improving factuality of LLMs without using external knowledge, including inference-

time intervention (Li et al., 2023b; Chuang et al., 2023), cross-exam (Cohen et al., 2023; Du et al., 2023), self-verify (Dhuliawala et al., 2023), or reinforcement learning (Tian et al., 2024; Wu et al., 2023). Our work differs from them in providing citations to external knowledge in the responses. Additionally, there is past work that also uses external knowledge (e.g., knowledge base) to reduce hallucination by injecting knowledge into prompts (Elaraby et al., 2023; Peng et al., 2023). While the external knowledge used for generating a response can possibly serve as a coarse and general reference, these approaches also do not offer granular, sentence-level citations as in our work.

Lastly, the proposed framework is a form of a retrieval augmented generation approach. While past work has explored using retrieval to improve LLM generation quality (Chen et al., 2017; Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2020; Shi et al., 2023) or factuality (Shuster et al., 2021; Jiang et al., 2023; Pan et al., 2023), our approach further enables LLMs to generate citations and self-generated citations to guide retrieval.

3 Problem & Background

Our proposed framework aims to adapt a pre-trained LLM $\mathcal{M}^{\mathcal{B}}$ to $\mathcal{M}^{\mathcal{A}}$ that is able to provide grounded responses with citations. Given a text query Q and a corpus $\mathcal{D} = \{d_i\}$ consisting of text passages, the adapted LLM $\mathcal{M}^{\mathcal{A}}$ is required to generate a response A to the query that is factually grounded in the corpus $\mathcal{D}$ as well as providing citations $\mathcal{C}$ together with its response.

Following past work (Liu et al., 2023; Gao et al., 2023b), we segment LLMs' output into statements by sentences and require each of the sentences to cite a set of passages from the corpus. Specifically, let $s_1, \ldots, s_n$ be the statements in the answer $A = s_1, \ldots, s_n$. The citations $\mathcal{C} = \{E_1, \ldots, E_n\}$ links each statement s_i to a set of evidence passages $E_i \subset \mathcal{D}$.

Recall that our adaptation aims to provide better grounded responses. With citations $\mathcal{C}$, we can quantify the grounding quality of a response A by a grounding score $\mathcal{G}$:

$$\mathcal{G}(A, \mathcal{C}) = \frac{1}{n} \sum_i \phi(\text{concat}(E_i), s_i),$$

where ϕ is an NLI model that assesses whether the concatenated passage $\text{concat}(\text{E}_\text{i})$ supports s_i. The grounding score $\mathcal{G}$ essentially averages how well each sentence is supported by its citations.

4 AGREE Framework

The proposed AGREE framework takes a holistic perspective for grounding, proposing a model tuning approach that adapts the base LLM to include citations in its responses, and introducing a test-time adaptation (TTA) mechanism that leverages the citation information for actively retrieving from the corpus and iteratively refining the responses.

4.1 Tuning LLMs

We tune the LLM to self-ground the claims in their responses by providing citations to retrieved documents. Our method is able to grant LLMs such an ability using only a collection of *unlabeled* queries $\{Q\}$ and an NLI model ϕ. As we are using unlabeled queries without reference responses, we formulate the adaptation task as tuning LLMs to achieve better grounding without heavily deviating from the original generations (such an approach of preservation has also been adopted in recent work (Gao et al., 2023a)). Conceptually, we adapt $\mathcal{M}^{\mathcal{B}}$ to $\mathcal{M}^{\mathcal{A}}$ so that the answers generated by the adapted LLM $\mathcal{M}^{\mathcal{A}}$ should satisfy the grounding constraints (with grounding score $> \tau_{\mathcal{G}}$) while maximizing the scores with respect to the base LLM $\mathcal{M}^{\mathcal{B}}$:

$$\max \mathbb{E}_{(A,\mathcal{C}) \sim \mathcal{M}^{\mathcal{A}}(\cdot|Q,\mathcal{D})} \mathcal{M}^{\mathcal{B}}(A \mid Q, \mathcal{D}) \mathbb{1}\{\mathcal{G}(A, \mathcal{C}) \geq \tau_{\mathcal{G}}\}. \quad (1)$$

In practice, we adopt a data-centric approach for optimizing $\mathcal{M}^{\mathcal{A}}$. For a given question, we opt to **use the maximally-grounded response sampled from the base LLM** to construct the tuning data. We will detail the process in the following of this section.

Data generation As shown in Fig 2, given the query, we first sample responses $\{A\}$ from the base LLM $\mathcal{M}^{\mathcal{B}}(\cdot \mid Q, \mathcal{D})$ using instruction following (see Appendix A for details). For each $A = s_1, \ldots, s_n$ we create citations $\mathcal{C} = \{E_i\}$ using the NLI model, ϕ, to link a sentence s_i to the maximally supported passage $e_i = \max_{d \in \mathcal{D}} \phi(d, s_i)$ if the passage e_i actually support s_i (i.e., $\phi(e_i, s_i) > \tau$).[1] Otherwise, we do not add a citation to s_i, and s_i is an unsupported statement. That is: $E_i = \{e_i\}\, if\, \phi(e_i, s_i) > \tau\, else\, \{\}$. We use U to

[1] In practice, we only present 5 passages retrieved from $\mathcal{D}$ to the LLM for generating initial responses, and only generate citations to this set of retrieved passages. We use TRUE (Honovich et al., 2022), a T5-11B NLI model. Please refer to Appendix A for more details.

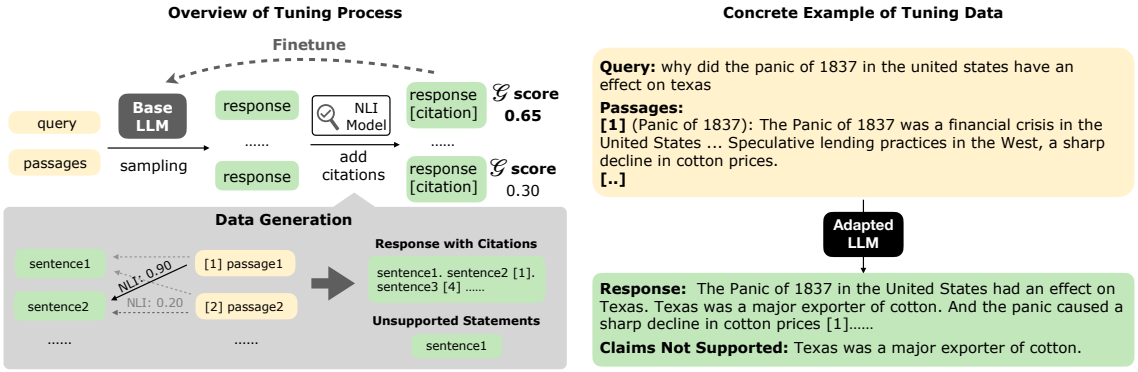

Figure 2: Illustration of the tuning process. We sample responses from the base model, use an NLI model to add citations to the sampled responses, and tune the base model with the best-grounded response. We also show a concrete example of tuning data on the right.

denote the set of unsupported statements that cannot find citations. This allows us to evaluate the grounding of A as in Section 3. Now, we can choose the best response A^* from $\{A\}$ based on the grounding scores to form a grounded response, i.e., $A^* = \arg\max_A \mathcal{G}(A, C)$.

We then use $\{Q, A^*, C^*\}$ (C^* as the citations associated with A^*) to teach the base LLM to generate grounded responses with citations. In addition to citations, we also instruct the LLM to clearly state the unsupported statements U^*, as shown in Fig. 2. We note that the tuning of framework does not force all training responses to be perfectly grounded. Instead, we supervise the LLM itself to identify unsupported statements. This allows the LLM to generate more flexibly and guide the retrieval process with its knowledge.[2]

Supervised fine-tuning We have introduced how we construct supervision to instruct the LLM to add citations and state unsupported statements in its response. To effectively tune the LLM, we verbalize the entire process in natural language. We denote the verbalized natural language description as $\text{VERB}(A^*, \mathcal{C}^*, U^*)$ (see Fig. 2 for a concrete example).[3] The natural language formalization also allows us to conveniently tune the LLM with standard language modeling objectives:

$$\mathcal{M}^{\mathcal{A}} = \arg\max_{\mathcal{M}} \sum_{Q} \mathcal{M}(\text{VERB}(A^*, \mathcal{C}^*, U^*) \mid Q, \mathcal{D}). \tag{2}$$

We note that this actual objective, Eq (2), maximizes the log probability of generating the best-grounded answer A^* that is selected from the generation of the base model. As A^* is sampled from the base model, such an objective avoids significant deviations from the original generations, which aligns with the goal of the conceptual objective (Eq (1)).

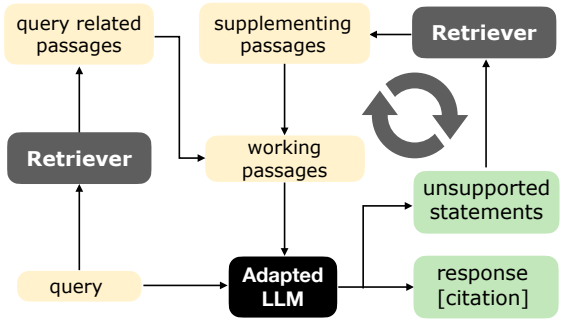

Figure 3: Illustration of the test-time adaptation mechanism. The adapted LLM retrieves from the corpus based on self-generated citation information to refine its response in an iterative way.

Multi-dataset training We use multiple existing datasets to construct the adaptation data used to tune the pre-trained LLM, including Natural Questions (NQ) (Kwiatkowski et al., 2019), FEVER (Thorne et al., 2018), and StrategyQA (Geva et al., 2021). We choose these as they contain diverse text, and the answers to the corresponding queries require different types of reasoning processes: NQ provides diverse queries naturally asked by real human users; FEVER places a particular emphasis on fact verification; and StrategyQA requires multi-hop reasoning with implicit strategy. It is worthwhile to note that AGREE *only* uses queries, leaving out ground-truth answers, to improve LLMs.

[2]Please refer to Appendix A for more details on the tuning method.

[3]Please refer to Appendix E for more examples of tuning data.

Algorithm 1 Iterative TTA

1: **procedure** ITERATIVEINFERENCE(Q, $\mathcal{D}$, $\mathcal{M}^{\mathcal{A}}$, k, B)
input: A query Q, text corpus $\mathcal{D}$, the adapted LLM $\mathcal{M}^{\mathcal{A}}$, the number of passages k that $\mathcal{M}^{\mathcal{A}}$ can take as input, the budget for LLM calls B
2: $relevant\_psgs = []$
3: ▷ retrieve passages using the query
4: $working\_psgs :=$ RETRIEVE$(Q, \mathcal{D})[: k]$
5: ▷ keep track of seen passages to avoid presenting duplicate passages to the LLM
6: $seen\_psgs := []$
7: **while** $iter = 1 : B$ **do**
8: ▷ Use the LLM to generate an answer A for the query Q based on the working psgs $\mathcal{D}$. Additionally obtain the cited passages and unsupported the sentences.
9: $A, cited\_psgs, unsup\_sents := \mathcal{M}^{\mathcal{A}}(Q, working\_psgs)$
10: ▷ add cited passages to the list of relevant passages and de-duplicate the list
11: $relevant\_psgs :=$ DEDUPLICATE$(relevant_psgs + cited_psgs)$
12: ▷ update the seen passages to include the working passages of this iteration
13: $seen\_psgs := seen\_psgs + working\_psgs$
14: **if** $unsup\_sents$ is not None **then**
15: ▷ retrieve additional information related to the unsupported statements
16: $supplementing\_psgs :=$ RETRIEVE$(unsup_sents, \mathcal{D})$
17: **else**
18: ▷ include more query-related passages to acquire more complete information
19: $supplementing\_psgs :=$ RETRIEVE$(Q, \mathcal{D})$
20: ▷ update the working passage to include supplementing passages that have not been presented to the LLM before
21: $working\_psgs :=$ DEDUPLICATE$(relevant_psgs +$ SETDIFF$(supplementing_psgs, seen_psgs))[: k]$
22: **return** $A, cited\_psgs$

4.2 Test-time adaptation

We introduce a novel test-time adaptation (TTA) method for the inference procedure, overviewed in Fig. 3. Our framework is a form of retrieval augmented generation framework – at the core of our approach lies the adapted LLM that is able to answer a query based on a set of given passages retrieved from the corpus, and, more importantly, self-ground its response to add citations to the passages as well as to find unsupported statements needing further investigation. With these capabilities, the adapted LLM can iteratively construct a set of relevant passages from the large corpus $\mathcal{D}$ and refine its response to the query.

The detailed procedure of TTA is shown in Algorithm (1). Given a query Q and the corpus $\mathcal{D}$, we first retrieve based on the query to obtain an initial set of working passages. Next, we employ the following procedure iteratively until we consume all the budget B of invoking LLM calls. At each iteration, the LLM generates a response to the query based on the working passages, adds citations to its response, and finds out any unsupported statements that do not have citations (ln 9). Then, we add the cited passages to the list of relevant passages. Lastly, at each iteration, we update the working passages – if there are unsupported statements, we include additional information retrieved based on the unsupported statements (ln 15), otherwise, we include more passages that are retrieved based on the query to acquire more complete information (ln 17). We only include passages that are new and haven't been presented to the LLM yet (ln 19). Note that at each iteration, we let the LLM to re-generate a response based on the current working passages instead of editing from previous one, which we observed lead to better fluency.

Dataset	Type	Corpus	#
	Train		
NQ	Factoid QA	Wiki	2500
StrategyQA	Multi-htop QA	Wiki	1000
Fever	Fact Checking	Wiki	1000
	In-Distribution Test		
NQ	Factoid QA	Wiki	700
StrategyQA	Factoid QA	Wiki	460
	Out-of-Distribution Test		
ASQA	Ambiguous QA	Wiki	948
QAMPARI	Multi-answer QA	Wiki	1000
Enterprise	Customer Support QA	Enterprise	580

Table 1: Statistics used for adaptation and test datasets. In addition to in-domain test datasets, we also investigate the generalization to out-of-distribution datasets that exhibit different reasoning processes or different corpus types.

The design of our proposed TTA enables efficient and flexible inference. We rely on the LLM to generate citations itself, which has the advantage of reduced overhead of invoking an additional NLI model in a post-hoc way. Also, as we iteratively refine the answer, such a process can be streamed and flexibly controlled by setting a budget in deployment.

5 Experiments

5.1 Setup

Evaluation datasets We conduct comprehensive evaluation on 5 datasets. Recall that we train AGREE on multiple datasets including NQ, Strate-

	NQ			StrategyQA			ASQA			QAMPARI			Enterprise	
	em-rec	rec	pre	acc	rec	pre	em-rec	rec	pre	rec-5	rec	pre	rec	pre
Base model: `text-bison-001`														
ICLCITE	47.6	52.1	56.3	74.5	13.6	27.8	39.5	47.3	49.8	20.3	22.7	24.5	30.2	40.5
POSTSEARCH	45.1	29.7	28.7	**75.5**	20.1	20.1	38.4	19.2	19.2	**22.5**	16.2	16.2	15.9	15.9
POSTATTR	45.1	31.5	31.5	75.5	18.4	18.4	35.1	38.0	38.0	22.5	18.5	18.5	20.1	20.1
AGREE$_{\text{w/o TTA}}$	50.0	67.9	73.1	74.1	33.4	50.5	39.5	65.9	70.5	20.1	60.1	64.5	55.8	67.1
AGREE$_{\text{w/ TTA}}$	**53.1**	**70.1**	**75.0**	74.9	**39.2**	**57.9**	**40.9**	**73.2**	**77.0**	20.9	**62.9**	**67.1**	**57.2**	**68.6**
Base model: `llama-2-13b`														
ICLCITE	45.8	42.8	41.6	**65.5**	20.6	33.1	35.2	38.2	39.4	**21.0**	10.2	10.4	30.6	38.8
POSTSEARCH	35.9	17.5	17.5	64.3	8.7	8.7	25.0	23.6	23.6	12.0	27.5	27.5	13.4	13.4
POSTATTR	35.9	26.0	26.0	64.3	12.5	12.5	25.0	33.6	33.6	12.0	28.9	28.9	18.7	18.7
AGREE$_{\text{w/o TTA}}$	47.9	50.5	56.6	65.0	25.5	35.0	35.7	50.2	55.3	17.1	40.4	43.6	**50.6**	53.8
AGREE$_{\text{w/ TTA}}$	**51.0**	**62.0**	**66.0**	64.6	**30.2**	**37.2**	**39.4**	**64.0**	**66.8**	17.9	**51.4**	**53.4**	50.4	**55.4**

Table 2: Answer accuracy and grounding (measured by citation quality) of AGREE and baselines across 5 datasets. Our approach achieves substantially better citation grounding (measured by citation recall) and citation precision compared to the baselines.

gyQA, and Fever. In addition to the two in-domain test sets, NQ and StrategyQA (we leave out the non-QA dataset, FEVER), we further test the generalization of adapted LLMs on 3 out-of-domain datasets, including ASQA (Stelmakh et al., 2022), QAMPARI (Amouyal et al., 2022), and an Enterprise dataset.[4] In particular, ASQA and QAMPARI contain questions of ambiguous answers and multiple answers. The Enterprise dataset is a proprietary dataset which requires provided answers that are grounded in customer service passages. Such an evaluation suite allows assessing the generalization capability of the adapted LLMs for OOD question types (ASQA and QAMPARI) as well as to an entirely different corpus (Enterprise).

Models We demonstrate AGREE framework with two LLMs, `text-bison` and `LlaMA-2-13B` (Touvron et al., 2023). We use GTR-large (Ni et al., 2021) as our retriever, and use TRUE (Honovich et al., 2022) as the NLI model.

Baselines We evaluate the effectiveness AGREE in two settings, invoking LLMs once, without TTA; and invoking LLMs multiple times, with the proposed TTA.[5] We compare with three baselines from recent work, including one prompting-based approach and two post-hoc citing approaches, described below.

Few-shot In-Context Learning (ICLCITE): Following Gao et al. (2023b), we prompt LLMs with few-shot examples (Gao et al., 2023b), each consisting of a query, a set of retrieved passages, and an answer with inline citations. The LLMs can therefore learn from the in-context examples and generated citations in the responses. It is worthwhile to note that ICLCITE *is a RAG baseline that also uses retrieved passages*.

Post-hoc search (POSTSEARCH): Following Gao et al. (2023b), given a query, we first instruct LLMs to answer the query *without* passages, and then add citations in a post-hoc way via searching. We link each claim in the response to the most relevant passage retrieved from a set of query-related passages. This baseline only uses the retriever but not the NLI model.

Post-hoc Attribution (POSTATTR): Following Gao et al. (2023a), instead of citing the most relevant passage, for each claim, we retrieve a set of k passages from the corpus, and then use the NLI model, ϕ, to link to the passage that maximally supports the claim. We note both baselines in the post-hoc citing paradigm only rely on LLMs' parametric knowledge.[6]

Metrics We mainly focus on improving the grounding quality of generated responses, reflected by the quality of citations. Following past work (Gao et al., 2023b), we report the **citation recall** (rec) and **citation precision** (pre) on all the evaluation datasets. We note that **citation recall**

[4] We use FEVER to create tuning data, but do not use it for evaluations. As we use LLMs in a zero-shot setting, the LLMs do not always answer with the specific labels defined in FEVER, which might introduce inaccuracies in the evaluation of answer correctness.

[5] We set the budget B for LLM calls used in TTA to be 4.

[6] Please refer to Appendix B for more details on experimental setup.

aggregates how well each sentence is supported by the citation to the corpus, which is essentially the grounding score $\mathcal{G}$. Therefore, we prioritize on the evaluation of citation recall.

We also report the correctness of the generated outputs. For NQ, we report exact match recall (em-rec; whether the short answers are substrings in the response). For StrategyQA, we report the accuracy (acc). For ASQA and QAMPARI, we use subsets from Gao et al. (2023b), and report the exact match recall (em-rec) for ASQA and recall-5 (rec-5, considering recall to be 100% if the prediction includes at least 5 correct answers) for QAMPARI. For the Enterprise dataset, we only report the citation quality as there are no ground truth answers for this dataset, and citation quality reflects whether the model can provide accurate information.

5.2 Results and analyses

Tuning is effective for superior grounding: Table 2 summarizes the results obtained using our AGREE framework and compares with the baselines. As suggested by the results, across 5 datasets, AGREE can generate responses that are better grounded in the text corpus and provide accurate citations to its response, substantially outperforming all the baselines. When tuned with high-quality data, LLMs can effectively learn to self-ground their response without needing an additional NLI model. On the other hand, ICLCITE, which solely relies on in-context learning, cannot generate citations as accurately as a tuned LLM, as suggested by the large gap on citation precision between ICLCITE and AGREE. We also observe similar findings as suggested by Gao et al. (2023b): POSTCITE often leads to poor citation quality – without being conditioned on passages, the response from POSTCITE often cannot be paired with passages that lead to high citation recall for the generated claims.

The performance improvements can generalize: Recall that we adapt the base LLM only using in-domain training sets (NQ, StrategyQA, and FEVER), and directly test the model on out-of-distribution (OOD) test set (ASQA, QAMPARI, Enterprise). The results suggest that the improvements obtained from training on in-domain datasets can effectively generalize to OOD datasets that contain different question types or use different types of corpus. This is a fundamental advantage of the proposed approach – AGREE can generalize to a target domain in the zero-shot setting without needing any samples from the target domain, which is needed for ICLCITE.

TTA improves both grounding and answer correctness: The comparison between AGREE without and with TTA highlights the effectiveness of the proposed iterative TTA strategy. We observe improvements in terms of both better grounding and accuracy. For instance, TTA improves `llama-2` answer correctness by 3.1 and 3.7 on NQ and ASQA, respectively. Such improvements can be attributed to the fact that our TTA allows the LLMs to actively collect relevant passages to construct better answers following the self-grounding guidance.

Discussions on answer correctness: In general, $\text{AGREE}_{\text{w/ TTA}}$ can achieve better correctness compared to ICLCITE. $\text{AGREE}_{\text{w/o TTA}}$ achieves similar answer correctness with ICLCITE, as both methods are conditioned on the same set of passages. As a result, the quality of passages heavily intervenes on the correctness of the answers. Unlike AGREE and ICLCITE, POSTATTR purely relies on the parametric knowledge of the LLMs to answer the query. As a result, POSTATTR generally achieves inferior answer correctness compared to AGREE and ICLCITE on these two LLMs, especially on the less capable LLM, `llama-2-13b`, that has less accurate knowledge compared to `bison`. Moreover, on the Enterprise dataset which contains more domain-specific information, POSTATTR utterly fails to recall attributable information from LLMs' parametric knowledge.

Results with different LLMs: Our approach successfully adapts both `text-bison-001` and `llama-2-13b`. `llama` is generally less capable compared to `bison`, underperforming `bison` in terms of answer correctness and citation quality. Still, AGREE also consistently outperforms the baseline, generating more grounded answers as well as providing more precise citations. This highlights that the proposed tuning-based adaptation approach is model-agnostic and is effective across LLMs of varying capabilities.

Computational efficiency: AGREE framework fine-tunes the base LLM to enable self-grounding without needing for additional in-context examples or NLI models. As a result, our framework is able to achieve strong citation performance without expensive inference cost. Table 4 shows the comparison between the computation cost, measured by the

	NQ			StrategyQA			ASQA			QAMPARI			Enterprise	
	em-rec	rec	pre	acc	rec	pre	em-rec	rec	pre	rec-5	rec	pre	rec	pre
Base model: `text-bison-001`														
ICLCITE	47.6	52.1	56.3	**74.5**	13.6	27.8	39.5	47.3	49.8	20.3	22.7	24.5	30.2	40.5
AGREE$^{\text{Multi-dataset}}_{\text{w/o TTA}}$	**50.0**	**67.9**	**73.1**	74.1	**33.4**	**50.5**	**39.5**	**65.9**	**70.5**	**20.1**	**60.1**	**64.5**	**55.8**	**67.1**
AGREE$^{\text{NQ-only}}_{\text{w/o TTA}}$	49.4	62.3	69.1	74.1	33.0	45.5	38.4	56.0	64.5	19.1	43.7	49.5	40.5	59.2
Base model: `llama-2-13b`														
ICLCITE	45.8	42.8	41.6	65.5	20.6	33.1	35.2	38.2	39.4	21.0	10.2	10.4	30.6	38.8
AGREE$^{\text{Multi-dataset}}_{\text{w/o TTA}}$	47.9	50.5	56.6	65.0	25.5	35.0	35.7	50.2	55.3	17.1	40.4	43.6	50.6	52.8
AGREE$^{\text{NQ-only}}_{\text{w/o TTA}}$	48.1	47.4	53.6	62.1	25.0	30.2	35.0	44.0	51.2	15.7	33.1	38.0	44.7	49.2
AGREE$^{\text{Distill}}_{\text{w/o TTA}}$	47.9	59.1	65.1	64.4	30.5	41.1	35.2	58.5	65.2	17.9	52.5	52.7	48.1	55.9

Table 3: Analysis on the impact of training data. Training with multiple datasets (AGREE$^{\text{Multi-dataset}}$) leads to better grounding (citation recall) and better citation precision across datasets, compared to training using the NQ dataset (AGREE$^{\text{NQ-only}}$). The citation quality of a less capable model `llama-2-13b` can also benefit from tuning using outputs from a more capable model (`text-bison-001`).

	# Tok: LLM	# Tok: NLI (T5-11B)
ICLCITE	2800	–
POSTATTR	360	3520
AGREE$_{\text{w/o TTA}}$	1210	–
AGREE$_{\text{w/ TTA}}$	4840	–

Table 4: The average computation cost (for one query) of different methods measured by the number of tokens processed by the LLM and the NLI model (based on a T5-11B architecture). AGREE$_{\text{w/o TTA}}$ is able to achieve better citation quality compared to ICLCITE, despite consuming less than half of the tokens needed for ICLCITE.

number of tokens processed by the LLM and the NLI model, needed for one query of our methods and that of the baselines. Compared to ICLCITE, AGREE$_{\text{w/o TTA}}$ uses much fewer tokens due to not using additional in-context examples, but achieves significantly better citation quality (see Table 2). POSTATTR does not use retrieved passages in the prompts and hence requires less computation on the LLM compared to our framework, but it requires additional overhead of extensively invoking the NLI model (which has 11B parameters – see Appendix A for details) to verify the each of the claims based on each of the retrieved passages. The citation performance of POSTATTR also substantially lags ICLCITE and AGREE. AGREE$_{\text{w/ TTA}}$ requires more computation compared to AGREE$_{\text{w/o TTA}}$, but is able to achieve both better citation quality and improvements in answer correctness.

The impact of training with multiple datasets: AGREE uses multiple datasets spanning factoid QA, multi-hop reasoning, and fact-checking to construct data for adapting the base model. We expect such a combination can grant the adapted model better generalization to different types of questions and different text distributions. We conduct an analysis to investigate the benefits of using multiple datasets for tuning. Table 3 shows the performance of our approach trained using multi-datasets and a counterpart that is trained only on NQ data (AGREE$^{\text{NQ-only}}$). The results suggest that training using NQ leads to inferior citation quality compared to training on the combination of three datasets across all the datasets. The performance gap is especially significant on datasets other than NQ. Moreover, training only on NQ also leads to inferior answer correctness across all the datasets. Nevertheless, training only on NQ can still improve performance compared to solely relying on in-context learning (ICLCITE).

AGREE as a distillation approach: Our work mainly focuses on improving the base LLM's grounding capability in a self-improving way. That is, we use the samples generated by the base LLM itself to adapt the base LLM, as opposed to distilling from proprietary models which may impose constraints on the deployment of adapted model. Nevertheless, we conduct an analysis to investigate the effectiveness as a distillation approach in a data-centric way by using data generated from a more capable model to enhance the grounding data for a less capable model. We use the data generated by `text-bison-001` to tune `llama-2-13b`.[7] As shown in the last row of Table 3, AGREE$^{\text{Distill}}$ achieves better citation quality compared to AGREE on `llama-2-13b`, as it is trained on better grounded responses produced by a more capable model. However, `llama-2` fine-tuned on data generated by `bison` still can't level the performance on `bison` as constrained by the gap between the capabilities of these

[7] Please refer to Appendix B for details.

Example from NQ

Query: where did the world's largest recorded wave occur?

Without TTA

Passage [1]: (retrieved using the query)
(Title: 1958 Lituya Bay earthquake and megatsunami) Lituya Bay has a history of megatsunami events, but the 1958 event was the first for which sufficient data was captured...
AGREE$_{\text{w/o TTA}}$: The world's largest recorded wave occurred in Lituya Bay,Alaska. The data was captured in a 1958 megatsunami event [1].

With TTA

Passage [2]: (retrieved using prior unsupported statement)
(Title: Megatsunami) In 1958, a 7.8 strike-slip earthquake in southeast Alaska caused 90 million tonnes of rock and ice to drop into the deep water at the head of Lituya Bay. The block fell almost vertically and hit the water with sufficient force to ...
Passage [3]: (retrieved using prior unsupported statement)
(Title: Tsunami) ...Their existence was confirmed in 1958, when a giant land slide in Lituya Bay, Alaska, caused the highest wave ever recorded, which had a height of 524 metres....
AGREE$_{\text{w/ TTA}}$: The world's largest recorded wave occurred in Lituya Bay, Alaska [3]. The wave was caused by a giant landslide that occurred in 1958 [2]. The landslide displaced 90 million...[2].

Example from ASQA

Query: Who played the weasley brothers in harry potter?

Without TTA

Passage [1]: (retrieved using the query)
(Title: James and Oliver Phelps) James Andrew Eric and Oliver Martyn John Phelps are identical twin British actors, best known for playing Fred and George Weasley in the "Harry Potter" film series.
AGREE$_{\text{w/o TTA}}$: The Weasley brothers in Harry Potter were played by identical twin actors James and Oliver Phelps [1].

With TTA

Passage [2]: (additional passages retrieved using the query)
(Title: Chris Rankin) ... Chris Rankin Christopher William "Chris" Rankin is a New Zealand-born British actor who is best known for playing Percy Weasley in the "Harry Potter" film franchise.
AGREE$_{\text{w/ TTA}}$: James and Oliver Phelps played Fred and George Weasley in the Harry Potter film series [1]. Chris Rankin played Percy Weasley in the Harry Potter film franchise [2].

Figure 4: Output examples of the proposed AGREE framework with `text-bison-001` as the base model. TTA is able to improve the response by retrieving more relevant information to precisely support a statement (see top) or finding more passages to generate a more complete response (see bottom).

two LLMs.

Qualitative analyses: We qualitatively analyze the advantages of the proposed AGREE framework compared to ICLCITE, the strongest among the baselines. We observe that on both `text-bison-001` and `llama-2-13b`, ICLCITE achieves inferior citation quality due to failure in following the citation format (e.g., adding citations after the periods, violating the instructions), linking a statement to a relevant but un-attributable passage (as indicated by poor citation precision), and introducing more auxiliary information not mentioned in the retrieved passages (as indicated by citation recall). Our AGREE framework mitigates these issues by tuning on well-grounded responses certified by the NLI model. We also provide example outputs in Fig. 4 comparing the outputs of AGREE with and without proposed TTA and observe that TTA can help find more supporting passages by active retrieving using unsupported statements (top) or iteratively find more passages to construct a more complete response (bottom).

6 Conclusion

We introduce a novel framework, AGREE, that adapts LLM for improved grounding. The proposed framework tunes a pre-trained LLM to self-ground its response in retrieved passages using automatically collected data. The integrated capability for grounding their responses further enables the LLM to improve the responses at test time. Our evaluations across five datasets demonstrate the benefits of the proposed learning-based approach compared to approaches that solely rely on prompting or the parametric knowledge of LLMs.

7 Limitations and future work

AGREE employs an automated data creation that relies on an NLI model, instead of humans. Thus, the citation quality is dependent on the NLI model. As suggested in Gao et al. (2023b); Honovich et al. (2022), one issue might be favoring "fully support" and cannot effectively detect "partially support". Thus, the adapted LLMs may favor adding "fully support" citations. One solution is to curate a set of human-annotated citations for "partially support", which we defer to future work. Also, our evaluation follows prior work (Rashkin et al., 2023; Gao et al., 2023a) and uses the NLI model to quantify the grounding and citation quality. Therefore, our work can encounter the same issue as past work: grounding and citation quality evaluation is limited by the capabilities of the NLI model.

AGREE uses created grounded responses to LLMs via supervised fine-tuning, as we demonstrate it leads to strong empirical results. It is also possible to treat grounding as a preference and RLHF (Ouyang et al., 2022) to tune LLMs, which we leave to future work. AGREE tuning incurs additional cost that is a one-time requirement for adapting the LLM. Considering the substantial grounding improvements, we believe this would be acceptable for most applications, especially for those with high-reliability requirements. Future work can possibly explore training a separate universal improved grounding model beyond task-specific adaptation.

We have mainly considered open domain question answering datasets focusing on information seeking tasks in English. Generalization to other

long form generation tasks and other languages can be important future work directions.

Lastly, adding citations to LLM-generated responses in AGREE might carry a shared risk with related research – a seemingly plausible but incorrect citation could potentially make an unsupported statement more convincing to users.

Acknowledgments

Thanks to anonymous reviewers for their helpful feedback, as well as to Jinsung Yoon, Andreas Terzis, Yanfei Chen, Ankur Taly, Lucas Zhang, and Tina Pang for their help with various aspects of this work.

References

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2022. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

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

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Baindoor Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *ArXiv*, abs/2305.13281.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *ArXiv*, abs/2309.11495.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *ArXiv*, abs/2305.14325.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate

text with citations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *ArXiv*, abs/2307.02185.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering.

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. *ACM Comput. Surv.*, 55(12).

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Li-Yu Daisy Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *ArXiv*, abs/2305.06983.

Ehsan Kamalloo, Aref Jafari, Xinyu Crystina Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *ArXiv*, abs/2307.16883.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. ArXiv:2304.09848.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nathan McAleese. 2022. Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei 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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *ArXiv*, abs/2112.07899.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell,

Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, Toronto, Canada.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

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*, pages 1–64.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *ArXiv*, abs/2301.12652.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? a.k.a. will llms replace knowledge graphs? *ArXiv*, abs/2308.10168.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of ACL*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

A Details of Data Generation

Recall that we create the tuning data by first sampling responses from the base LLM and then using the NLI model to create citations and identify unsupported statements. We provide the details on the process in the following of this section.

NLI model We use TRUE NLI (Honovich et al., 2022) as, to the best of our knowledge, it is the state-of-the-art NLI model for evaluating whether a passage supports a claim, and it is commonly used in the recent line of work on attributed QA and evaluating grounding (Bohnet et al., 2022; He et al., 2022; Li et al., 2023a).[8] TRUE is trained on data collected from 6 datasets of diverse tasks covering NLI, paraphrase detection, and fact verification, which leads to its strong performance across diverse types of text. Furthermore, the citation performance evaluated by TRUE highly aligns with human evaluation (Gao et al., 2023b).

Corpus & retriever As mentioned before, our framework is an instantiation of retrieval-augmented framework. For the datasets using Wikipedia as the corpus (NQ, StrategyQA, ASQA, and Qampari), we use the 2018-12-20 Wikipedia snapshot as the corpus and set up the retriever using GTR-large (Ni et al., 2021).

> Task: You will be given a question and some search results. Please answer the question in 3-5 sentences, and make sure you mention relevant details in the search results. You may use the same words as the search results when appropriate. Note that some of the search results may not be relevant, so you are not required to use all the search results, but only relevant ones.
>
> <Question>
>
> Search Results:
> [<Index>] <Title>
> <Text>
>
> [...]
>
> Answer:

Figure 5: Zero-shot prompt template for sampling initial responses from the base LLM.

Sampling initial responses We sample initial responses from the base LLM using instruction following in a *zero-shot fashion*. Given a query, we present the base LLM with query and 5 retrieved passages appended after an instruction that requires the base LLM to answer the query based on the passages; see Fig. 5 for the template of the zero-shot prompt. We note that we opt to use a zero-shot prompt as opposed to a task-specific few-shot prompt since 1) this can avoid biasing the generation with the few-shot in-context examples, and 2) this matches the expected scenario for deploying the adapted LLM to handle new queries in a zero-shot fashion.

For `text-bison-001`, we sample 4 responses using a temperature of 0.5. For `llama-2-13b`, we sample 4 responses using nuclear sampling (Holtzman et al., 2019) with p=0.95.

Adding citations and identifying unsupported statements After obtaining the initial response $\{A\}$ from the base LLM. We break each response A into sentences into $s_1, \ldots, s_i$. For each s_i, we find the maximally supported passage e_i (scored by $\phi(e_i, s_i)$) that the base LLM has seen during generating the initial responses. We link e_i to s_i if $\phi(e_i, s_i) > 0.7$ to encourage more precise citations. For a sentence s_i if there does *not* exist an e_i such that $\phi(e_i, s_i) > 0.5$ (the decision boundary for entailment), we add s_i to the unsupported statement set U.

Verbalizing We show the template for verbalizing the data used to tune the LLM in Fig. 6. As shown in the figure, we verbalize the citations in enclosed box brackets that are added at the end of sentences (before periods) like [n], and verbalize unsupported statements after the responses.

B Details of Experimental Setup

Details of Finetuning For tuning, we use LORA tuning (Hu et al., 2022) in experiments on both `text-bison-001` and `llama-2-13b`. For `bison`, we use API to perform tuning.[9] and follow all the default hyper-parameters except for training steps. We set 10% data created as development data and choose to use a training step of 1000 (chosen from 500, 1000, and 2000). For `llama-2`, we use the huggingface transformers (Wolf et al., 2019) chat-version checkpoint.[10] We find the chat-version achieves better performance than the base checkpoint in our preliminary investigation. We set lora_r to be 32, and only choose to use a learning rate of 1e-5 (chosen from 1e-4 and 1e-5) using the development set.

[8] https://huggingface.co/google/t5_xxl_true_nli_mixture

[9] https://cloud.google.com/vertex-ai/docs/generative-ai/models/tune-text-models-supervised

[10] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

Input

Task: You will be given a question and some search results. You are required to perform the following steps.

First, please answer the question in 3-5 sentences, and make sure you mention relevant details in the search results. You may use the same words as the search results when appropriate. Note that some of the search results may not be relevant, so you are not required to use all the search results, but only relevant ones. If you use the provided search results in your answer, add [n]-style citations.

Next, review your response and find the unsupported sentences that do not have citations.

<Question>

Search Results:
[<Index>] <Title>
<Text>

[...]

Output

Answer: <Response with citations>

Sentences Not Supported by Citations: <Unsupported statements>

Figure 6: Verbalization template for creating the training data for adapting the base LLM.

We fine-tune `llama-2` on two A100 (40GB) GPU for 4 epochs.

Details of Decoding Our evaluation uses the official implementation from ALCE (Gao et al., 2023b), we use the same data split and prompt template from ALCE. We use temperature 0.25 for evaluation on both `bison` and `llama`. We use one sample for evaluation since adapted LLMs tend to generate better-grounded response exhibiting less variation.

Details of Distillation For distillation, we directly tune `llama-2-13b` using the data created with `text-bison-001`. We also set lora_r to be 32, use a learning rate of 1e-5, and fine-tune for 4 epochs.

C Comparison to ICLCITE on More Capable LLMs

Table 5 compares AGREE using `llama-2-13B` as the base model against ICLCITE on more capable models. We directly use the results from ALCE (Gao et al., 2023b). Our framework is able to substantially shorten the gap between a small llama-2 model and much more capable LLMs.

	ASQA			QAMPARI		
	em-rec	rec	pre	rec-5	rec	pre
	Base model: `llama-2-13b`					
AGREE$_{\text{w/o TTA}}$	35.7	50.2	55.3	17.1	40.4	43.6
AGREE$_{\text{w/ TTA}}$	39.4	64.0	66.8	17.9	**51.4**	**53.4**
	Base model: `llama-2-70b`					
ICLCITE	**41.5**	62.9	61.3	**21.8**	15.1	15.6
	Base model: `ChatGPT-0301`					
ICLCITE	40.4	**73.6**	**72.5**	20.8	20.5	20.9

Table 5: Comparing AGREE on `llama-2-13B` against ICLCITE on `llama-2-70B` and `ChatGPT-0301`. We directly quote results from ALCE.

D License of Datasets

The licenses datasets used in our work include:

- NQ (Kwiatkowski et al., 2019) under Creative Commons Share-Alike 3.0 license.
- StrategyQA (Geva et al., 2021) under MIT License.
- Fever (Thorne et al., 2018) under Creative Commons Share-Alike license.
- Ambiguous QA (Stelmakh et al., 2022) under Creative Commons Share-Alike 3.0 license.
- Qampari (Amouyal et al., 2022) under Creative Commons Zero v1.0 Universal license.

E Additional Examples of Tuning Data

Please see Fig. 7 and Fig. 8 for concrete examples of tuning data.

Input

Task: You will be given a question and some search results. You are required to perform the following steps.

First, please answer the question in 3-5 sentences, and make sure you mention relevant details in the search results. You may use the same words as the search results when appropriate. Note that some of the search results may not be relevant, so you are not required to use all the search results, but only relevant ones. If you use the provided search results in your answer, add [n]-style citations.

Next, review your response and find the unsupported sentences that do not have citations.

Question:
what kind of sentence contains an independent clause and a dependent clause?

Search Results:
[1] Dependent clause
... A "complex sentence" contains an independent clause and at least one dependent clause. A sentence with two or more independent clauses plus (one or more) dependent clauses is referred to as a "compound-complex sentence". (Every clause contains a subject and predicate.)...

[2] Sentence clause structure
... A complex sentence has one or more dependent clauses (also called subordinate clauses). Since a dependent clause cannot stand on its own as a sentence, complex sentences must also have at least one independent clause...

[3] Sentence clause structure
A set of words with no independent clause may be an "incomplete sentence", also called a "sentence fragment". A sentence consisting of at least one dependent clause and at least two independent clauses may be called a "complex-compound sentence" or "compound-complex sentence". Sentence 1 is an example of a simple sentence...

[4] Topic sentence
Using a complex sentence is a great way to refer to the content of the paragraph above (dependent clause) and then bring in the content of the new paragraph (the independent clause). Here is a typical example: The beginning, dependent, clause probably refers to the content of a preceding paragraph that presented the ant as a community-focused worker...

[5] Sentence clause structure
This simple sentence has one independent clause which contains one subject, "I", and one verb, "run". This simple sentence has one independent clause which contains one subject, "girl", and one predicate, "ran into her bedroom"...

Output

Answer:
A complex sentence contains an independent clause and at least one dependent clause [1]. A dependent clause cannot stand on its own as a sentence, so complex sentences must also have at least one independent clause [2].

Sentences Not Supported by Citations:
None.

Figure 7: An example of tuning data created from NQ.

Input

Task: You will be given a question and some search results. You are required to perform the following steps.

First, please answer the question in 3-5 sentences, and make sure you mention relevant details in the search results. You may use the same words as the search results when appropriate. Note that some of the search results may not be relevant, so you are not required to use all the search results, but only relevant ones. If you use the provided search results in your answer, add [n]-style citations.

Next, review your response and find the unsupported sentences that do not have citations.

Question:
Could Robert Wadlow hypothetically see Frankenstein's monster's bald spot from above?

Search Results:
[1] Frankenstein's monster
In this version, Frankenstein gives the monster the brain of his mentor, Doctor Waldman, while his body is made from a man who killed Waldman while resisting a vaccination...In the 2004 film "Van Helsing ," the monster is shown in a modernized version of the Karloff design. He is tall, has a square bald head, gruesome scars, and pale green skin...

[2] Rorschach (character)
Rorschach is 5'6"tall and weighs 140 pounds, and, as Walter Kovacs (his"disguise"), he appears as a red-haired, expressionless, man who always carries with him a sign that reads "THE END IS NIGH"...

[3] Glenn Strange
During the wave of monster-related merchandising in the late 1950s and 1960s, Glenn Strange's iconic image often was used for the monster on toys, games, and paraphernalia, most often from his appearance in the Abbott and Costello film...

[4] John Bloom (actor)
John Bloom was an American actor. Career. Standing at 7' 4"(224 cm) tall, he is most famous for his roles as the Frankenstein Monster in the low-budget horror movie" Dracula vs. Frankenstein "(he remains the tallest actor to ever portray the character)"s...

[5] Model figure
There were also cheap plastic models for the popular market of movie monsters, comic book heroes, and movie and television characters in 1:8 size (about 9 inches or 23 cm in height)...

Output

Answer:
Yes, Robert Wadlow could hypothetically see Frankenstein's monster's bald spot from above. Frankenstein's monster is described as having a square bald head [1]. Robert Wadlow was 8 feet 11 inches tall. Frankenstein's monster was portrayed by John Bloom, who was 7 feet 4 inches tall [4].

Sentences Not Supported by Citations:
Robert Wadlow was 8 feet 11 inches tall.

Figure 8: An example of tuning data created from StrategyQA.