Learning to Generate Answers with Citations via Factual Consistency Models

Rami Aly[1]*, **Zhiqiang Tang**[2], **Samson Tan**[2], **George Karypis**[2]
[1]University of Cambridge [2]Amazon Web Services
rami.aly@cl.cam.ac.uk, {zqtang|samson|karypis}@amazon.com

Abstract

Large Language Models (LLMs) frequently hallucinate, impeding their reliability in mission-critical situations. One approach to address this issue is to provide citations to relevant sources alongside generated content, enhancing the verifiability of generations. However, citing passages accurately in answers remains a substantial challenge. This paper proposes a weakly-supervised fine-tuning method leveraging factual consistency models (FCMs). Our approach alternates between generating texts with citations and supervised fine-tuning with FCM-filtered citation data. Focused learning is integrated into the objective, directing the fine-tuning process to emphasise the factual unit tokens, as measured by an FCM. Results on the ALCE few-shot citation benchmark with various instruction-tuned LLMs demonstrate superior performance compared to in-context learning, vanilla supervised fine-tuning, and state-of-the-art methods, with an average improvement of 34.1, 15.5, and 10.5 citation F_1 points, respectively. Moreover, in a domain transfer setting we show that the obtained citation generation ability robustly transfers to unseen datasets. Notably, our citation improvements contribute to the lowest factual error rate across baselines.

1 Introduction

Large language models (LLMs; Brown et al., 2020) have showcased remarkable capabilities in information-seeking tasks. However, concerns persist regarding their trustworthiness, stemming from their opaque nature and factual hallucinations in their responses (Zhang et al., 2023; Mallen et al., 2023). One proposed solution involves the production of citations to retrieved text passages alongside LLM-produced statements (Liu et al., 2023; Rashkin et al., 2023). By integrating citations into generated responses, readers can easily verify LLMs statements. The ability to accurately produce citations enables LLMs to generate responses more closely aligned with cited sources, alleviating hallucinations (Gao et al., 2023b; Yue et al., 2023).

Question: Who has the highest goals in world football?

Retrieved Passages:
[1] ...Pelé's 1281 goals are recognized by FIFA as the highest total achieved by a professional footballer, although the Soccer Statistic Foundation (rssf) recognizes only 767 goals in official mode, occupying the third place after Josef Bican (805) and Romario (772) ...
[2] have beaten Gerd Müller's record of 85 goals in a year, the Football Association of Zambia ...

Answer (LlaMA-2-Chat-70B): According to IFFHS International Federation of Football History and Statistics, Lionel Messi has the highest goals in world football with 1281 goals **[1]**. However, some sources, such as the Soccer Statistic Foundation (rssf), recognize only 767 goals in official mode, placing Messi in third place after Josef Bican (805) and Romario (772) **[2]**.

Figure 1: An ALCE-ASQA question with a generated answer prompted via in-context learning. Two error classes are common: information not supported by the sources (red) and incorrect citation to the sources (blue).

Despite its significance, accurate citation generation proves to be challenging. State-of-the-art LLMs, such as ChatGPT (OpenAI, 2023), and commercial generative chat engines, such as Bing Chat, produce accurate citations only for less than 60% of generated statements (Gao et al., 2023b; Liu et al., 2023). Figure 1 illustrates typical citation errors, including hallucinated statements or citations associated with incorrect claims. Hence, there is a necessity to train LLMs to generate citations accurately. This paper focuses on teaching LLMs to generate citations for retrieval-augmented long-form question answering (LFQA), tackling two main challenges: the scarcity of high-quality labeled data at scale and the risk of compromising original language and generalization capacities during fine-tuning for citation generation.

* Work done while interning at Amazon Web Services.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11876–11896

To address these challenges, we present **CaLF** (**Ca**l**i**ta**t**ion **L**earning via **F**actual Consistency Models), a fine-tuning strategy that enables LLMs to learn citation generation without sacrificing their language capacities. As illustrated in Figure 2, the cornerstone of our approach is factual consistency models (FCMs; Kryscinski et al., 2020, *inter alia*), initially introduced as a neural measure of consistency between a claim and its context. We use FCMs to gauge whether cited passages support a generated statement. Our method incorporates FCMs in two designs. Firstly, we propose a weakly-supervised training strategy, where an LLM generates diverse responses with citations, an FCM filters high-quality citation data, and the LLM is fine-tuned on the filtered data. Secondly, we utilize focused learning to adjust the loss contribution of each answer token based on its factual relevance, as measured by an FCM. The intuition is to have the LLM concentrate on tokens related to factual knowledge during fine-tuning, minimizing the impact on its original language capacities.

We evaluate CaLF on various LLMs, including Llama2 (Touvron et al., 2023), Mistral-Instruct, and MistralOrca (Jiang et al., 2023). On the ALCE automatic few-shot citation evaluation benchmark (Gao et al., 2023b), CaLF enhances citation metrics over the in-context learning and baseline fine-tuning, with an average improvement of 34.1 and 15.5 F_1, respectively, while maintaining fluency and correctness. All LLMs trained via CaLF , outperform the state-of-the-art model Self-Rag (Asai et al., 2024) and ChatGPT (OpenAI, 2023), with an average improvement of 24.8 and 10.5 citation F_1 points, respectively. Domain transfer experiments, testing citation quality on a dataset different from the training dataset, highlight CaLF's ability to generalize across tasks and domains. Additionally, on the FactScore biography generation benchmark (Min et al., 2023), CaLF demonstrates an improvement in factuality. Finally, human evaluation results indicate that CaLF yields more preferable answers compared to the fine-tuning baseline.

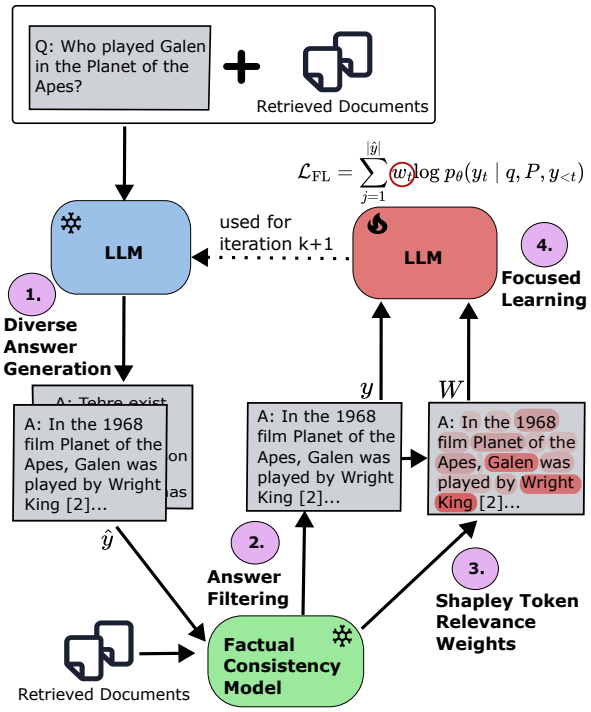

Figure 2: A schematic view of our iterative citation fine-tuning method CaLF. It uses a factual consistency model to: i) create weakly supervised training instances by filtering diversely sampled responses, ii) adjust the loss contribution of each answer token according to its Shapley relevance for factual consistency prediction.

2 Related Work

LFQA with Citations To produce citations alongside a response, the generation can be conditioned on a few high-quality in-context examples with embedded citations (Gao et al., 2023b; Li et al., 2023). In contrast, Gao et al. (2023a); Bohnet et al. (2022) propose to edit an already generated response post-hoc to attribute to retrieved sources, causing computational overhead during inference. Alternatively, proprietary work has explored human preference learning for citation production (Menick et al., 2022; Thoppilan et al., 2022; Nakano et al., 2021). Reinforcement learning from human feedback is expensive and typically more brittle than supervised fine-tuning. Very recently, Asai et al. (2024) incorporate critique tokens, which serve as a feedback and citation mechanism, into GPT-4 obtained instruction-tuning data. These tokens allow for flexible retrieval-augmented generation with citations. In contrast to previous work, CaLF incorporates openly available FCMs as critique models into the fine-tuning process of already instruction-tuned LLMs while not modifying inference, maintaining efficiency.

Factual Consistency Models FCMs assess whether all the factual information in a claim is consistent w.r.t to the information conveyed in its grounding text. The task shares strong similarities to natural language inference (NLI) (Bowman et al., 2015; Dagan et al., 2005). However, in contrast to NLI, factual consistency is not evaluated on subjec-

tive or opinionated statements. Work that explores FCMs for *improving* the generation is mainly constrained to summarization. Aharoni et al. (2023) filter samples from a large-scale dataset according to an FCM while Muller et al. (2023) use an FCM to rerank and select source passages for cross-lingual summarization. Tian et al. (2023) incorporate FCMs to improve factuality via direct preference optimization (Rafailov et al., 2023). Instead of using FCMs, Deng et al. (2023) improve the factuality of generations by measuring the cosine similarity between a token's embedding and relevant knowledge to adjust a token's loss contribution. Our work is the first to explore the use of model explainability mechanisms, namely Shapley values (Shapley, 1953), for incorporating FCMs directly into the fine-tuning process of an LLM.

3 Preliminaries

Task description. Given an information-seeking question q, such as shown in Figure 1, the task is to generate a long-form answer $\hat{y} = \{s_1, \cdots, s_n\}$, consisting of sentences s_i, conditioned on passages P retrieved from a knowledge base. A long-form answer *with citations* needs to ensure that one or multiple relevant passages $C_i \subseteq P$ are cited in each generated sentence s_i (indicated by brackets with a passage index, e.g. "[1]" for p_1), such that the generated information in s_i follows from the cited passages C_i. This task definition strictly requires all facts to originate from the retrieved passages, ensuring that $\hat{y}$ is fully verifiable by P. We further assume the availability of few-shot training samples $(q, P, y) \in \mathcal{D}$ to learn citation production.

Generation with LLMs. In this work, we are interested in using LLMs to generate long-form answers with citations. An answer $\hat{y}$ is generated autoregressively, computing the next token distributions conditioned on the question q, retrieved passages P, and the answer generated so far $\hat{y}_{<t}$: $\prod_{t=1}^{|\hat{y}|} p_\theta(\hat{y}_t \mid q, P, \hat{y}_{<t})$, with θ being the parameterization of the LLM. Consequently, a model is updated on a gold answer y by minimizing the negative log-likelihood loss (NLL):

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{\mid y \mid} \sum_{t=1}^{|y|} \log p_\theta(y_t \mid q, P, y_{<t}), \quad (1)$$

Factual Consistency Models. An FCM ϕ measures the factual consistency between a sentence s_i and a collection of citations C_i: $o = \phi(s_i, C_i)$, with $o \in \{0, 1\}$ being the binary consistency prediction of the FCM. While FCMs are often modelled via modified NLI models that output a binary prediction directly (Gekhman et al., 2023; Utama et al., 2022), AlignScore (Zha et al., 2023) produces a single calibrated continuous value $o = p_\phi(\text{consistent} \mid s_i, C_i) \in [0, 1]$. In our scenario, such a scalar is beneficial for computing Shapley values and for controlling the consistency's strictness. We refer to Mosca et al. (2022) for background on Shapley values and the SHAP framework in the context of NLP.

4 Citation Learning via FCMs

CaLF is a fine-tuning strategy for producing long-form answers with citations. CaLF's main assumption is that an FCM ϕ can be leveraged as a supervision signal to improve citation quality of an LLM $\mathcal{F}$ via an iterative training procedure, as illustrated in Figure 2 and Algorithm 1. CaLF alternates between two modes in each iteration k. First, it generates weakly-supervised training data $\hat{\mathcal{D}}_k$ to enrich the training corpus by sampling diverse answers from a fine-tuned LLM $\mathcal{F}_{k-1}$ and filters them via the FCM ϕ (Sec. 4.1). Second, it fine-tunes the LLM $\mathcal{F}_{k-1}$ on $\hat{\mathcal{D}}_k + \mathcal{D}$ with a modification of the NLL objective, which re-weights the loss contribution of individual tokens according to their importance for ensuring factual consistency, as measured by the FCM ϕ (Sec. 4.2). The number of iterations is determined dynamically by stopping when the proportion of filtered examples over the candidates does not improve between iterations or once the maximum number of iterations is reached.

4.1 Answer Generation for Training

To generate weakly supervised training data $(x, P, \hat{y}) \in \hat{\mathcal{D}}_k$ with answer $\hat{y}$ containing citations to passages P, we assume the availability of a collection of information-seeking questions $x \in \mathcal{X}$ and a list of atomic facts A expected in an answer to x. For questions $\mathcal{X}$, an LLM generates a collection of answer candidates $\hat{\mathcal{Y}}$, conditioned on retrieved passages P, selected by an out-of-the-box retrieval system $\mathcal{R}$. As indicated in Algorithm 1, we produce answer candidates using either the fine-tuned model F_{k-1} from the previous iteration, or, in the case $k = 0$, we use in-context prompting with the few-shot examples $\mathcal{D}$. We focus on the generation of *diverse answer candidates* to enrich the weakly supervised training data. First, we

Algorithm 1 The training procedure of CaLF.

Input: LLM $\mathcal{F}$; FCM ϕ; Retriever $\mathcal{R}$; questions $\mathcal{X}$ and answer facts $\mathcal{A}$; few-shot examples $\mathcal{D}$; Iterations K.
Output: Fine-tuned LLM $\mathcal{F}_K$; Citation Data $\hat{\mathcal{D}}_K$.

$\mathcal{U}_0 \leftarrow \{\oplus_{s_i \in y}(\text{Norm}(\text{SHAP}_\phi(s_i))) \mid (x, P, y) \in \mathcal{D}\}$
$\mathcal{W}_0 \leftarrow \{\text{Align}(U, \mathcal{T}_\phi(\hat{y}), \mathcal{T}_\mathcal{F}(\hat{y}) \mid U \in \mathcal{U}_0\}$
$\mathcal{P} \leftarrow \{\mathcal{R}(x, \text{KB}) \mid x \in \mathcal{X}\}$
$k \leftarrow 0$
while $k \leq K$ **and** $\frac{|\hat{D}_{k-1}|}{|\hat{Y}_{k-1}|} \geq |\frac{\hat{D}_{k-2}}{\hat{Y}_{k-1}}|$ **do**
  if $k = 0$ **then** ▷ Data Generation (§4.1)
    $\hat{Y}_k \leftarrow \text{Diverse Sampl.}_{\text{IC}}(\mathcal{F}, \mathcal{X}, P, \mathcal{D})$
  else
    $\hat{Y}_k \leftarrow \text{Diverse Sampl.}(\mathcal{F}_{k-1}, \mathcal{X}, P)$
  end if
  $\hat{D}_k \leftarrow \{(x, P, \hat{y}) \mid \hat{y} \in \hat{Y}_k \wedge \mathcal{Q}(\hat{y}, A, P) = 1\}$
  $\mathcal{U}_k \leftarrow \{\bigoplus_{s_i \in \hat{y}} \text{Norm}(\text{SHAP}_\phi(s_i))) \mid (x, P, \hat{y}) \in \hat{\mathcal{D}}_k\}$
  $\mathcal{W}_k \leftarrow \{\text{Align}(\mathcal{T}_\phi(\hat{y}), \mathcal{T}_\mathcal{F}(\hat{y}) \mid U \in \mathcal{U}_k\}$
  $\mathcal{F}_k \leftarrow$ Update $\mathcal{F}$ via FL ▷ Focused Learning (§4.2)
    $\nabla \mathcal{L}_{\text{FL}}(\mathcal{D} + \hat{\mathcal{D}}_k, \mathcal{W}_0 + \mathcal{W}_k)$
  $k \leftarrow k + 1$
end while

use sampling strategies such as nucleus sampling (Holtzman et al., 2020), temperature scaling (Guo et al., 2017), and diverse beam search (Vijayakumar et al., 2018). Second, we consider citation replacements in answer sentences s_i in $\hat{y}$ to diversify the answer candidates beyond the output of the LLM. Specifically, for generated citations C_i in a sentence s_i, we generate two citation replacements, sampled according to the passage probability measured via the retriever $\mathcal{R}$ since the direct computation of $\phi(s_i, C_i)$ over all citation options is infeasible.

Each answer candidate $\hat{y} \in \hat{\mathcal{Y}}$ is subsequently filtered via an answer quality assurance function $\mathcal{Q}_\phi : \hat{y} \rightarrow \{0, 1\}$, measured via the FCM ϕ, to obtain $\hat{\mathcal{D}}_t$ with weakly-supervised cited answers $\hat{y}$:

$$\mathcal{Q} = \begin{cases} 1 & \text{if} \quad \text{Citation-Recall}(\hat{y}, C, \phi) > \Theta \\ & \quad \wedge \text{Citation-Precision}(\hat{y}, C, \phi) > \Theta \\ & \quad \wedge \text{Correctness}(\hat{y}, A, \phi) > \Theta \\ 0 & \text{else,} \end{cases}$$

with C being the citations assigned to sentences in $\hat{y}$, and Θ being a dynamically determined quality threshold that adjusts such that the size $\hat{D}_t$ is above a minimum viable size. $\text{Citation-Recall}(\hat{y}, C) = \frac{1}{n} \sum_{s_i \in \hat{y}} \phi(s_i, C_i)$ measures the factual coverage of citations, $\text{Citation-Precision}(\hat{y}, C) = \frac{1}{|C|} \sum_{C_i \in C} \frac{1}{|C_i|} \sum_{c_{i,j} \in C_i} \max(\phi(s_i, c_{i,j}), 1 - \phi(C_i \backslash \{c_{i,j}\}, s_i))$ measures the relevance of citations, and $\text{Correctness}(\hat{y}, A)$ measures the proportion of facts A covered in $\hat{y}$. These definitions largely align with the ones in the benchmark of Gao et al. (2023b) for evaluating citations.

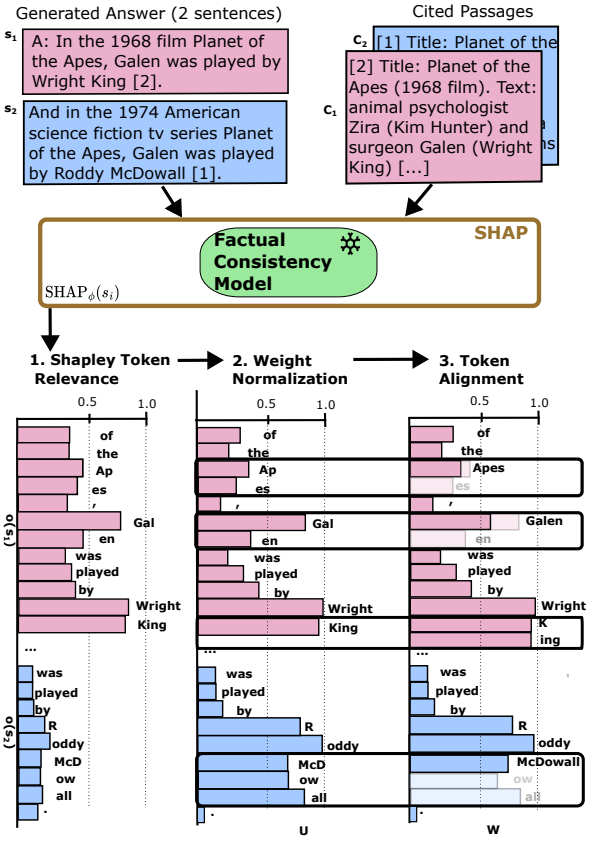

Figure 3: The computation of relevance weights W for rescaling the loss according to Eq. 2. We first use SHAP to measure the token importance for predicting $\phi(s_i, C_i) = o_i$. We adjust for differences in scale of $W_{\phi,i}$ for sentences s_i and differences in tokenization between the FCM and the LLM.

4.2 Focused Learning for Factual Consistency

To emphasize the learning of producing citations, we measure the relevance for each token in an answer for ensuring the factual consistency between y and the retrieved passages P. We subsequently modify the NLL loss computation (see Eq. 1) for the instruction-tuning of the LLM $\mathcal{F}$ by re-weighting the loss contribution of the $t-$th token according to relevance weights $w_t \in W$:

$$\mathcal{L}_{\text{FL}} = -\frac{1}{|\ y\ |} \sum_{j=1}^{|y|} w_t \log p_\theta(y_t \mid q, P, y_{<t}). \quad (2)$$

In contrast to NLL where the loss is computed as the arithmetic mean over the token-level loss, our focused learning loss $\mathcal{L}_{\text{FL}}$ emphasizes tokens which are considered of higher importance for ensuring

factual consistency between a generated statement and the cited passages according to an FCM ϕ.

The mechanism for obtaining the relevance weights W is illustrated in Figure 3. We first compute and normalize Shapley values for the factual consistency prediction between y and its cited passages using the FCM ϕ, obtaining U. These normalized relevance values for the tokens of the FCM are subsequently mapped to tokens of the LLM $\mathcal{F}$ via an alignment algorithm. As seen in Figure 3, given a generated statement such as "*In the 1968 film Planet of the Apes, Galen was played by Wright King.*", the relevance weights W consider factual tokens such as *Wright* and *King* more important than *of* and *was*, emphasizing units of information important for factually accurate citation.

Computation of Token Relevance. We first compute relevance weights U over the FCM ϕ by computing Shapley values for the factual consistency prediction between an answer's sentence s_i and its citations C_i: $o_i = \phi(s_i, C_i)$. Shapley values assign importance scores w_t to each feature (here token) in s_i concerning prediction o_i. Since Shapley values distribute the prediction score o_i along all the tokens of sentence s_i, the value of o_i and the length of s_i impacts the scale of assigned values, as shown in Figure 3, potentially biasing the loss towards shorter sentences and amplifying idiosyncrasies of the FCM. Thus, we normalize the assigned Shapley values for each sentence via min-max normalization $\text{Norm}(U_i)$: $\{\frac{u_t - \min(U_i)}{\max(U_i) - \min(U_i)} \mid u_t \in U_i\}$, with U_i being token weights for sentence s_i. Thus, the token with the highest and lowest Shapley value is assigned a relevance score of 0 and 1, respectively.[1] The computation of weights U for a given response y can subsequently be summarized as:

$$U = \oplus_{i=1}^{n} \text{Norm}(\text{SHAP}_\phi(s_i)) \quad (3)$$

with $\oplus$ being a concatenation operator. Since citation tokens do not bear any semantic meaning themselves, they are excluded from the computation of $\phi(s_i, C_i)$ and are thus not yet contained in U. Therefore, we further insert a weight of 1 for each citation token of y into U.

Feature Importance Mapping. Since the obtained relevance weights U are specific to the tokenizer used for ϕ, an LLM that uses a different tokenization may not directly apply U to re-weight the loss $\mathcal{L}_{\text{FL}}$ (c.f. Eq. 2). To this end, we define an alignment function that maps the FCM token weights U to LLM token weights W for the same sequence y. The alignment function first maps the shortest possible token span $y_{\phi,l:l+m}$ from the FCM's tokenizer T_ϕ to the span $y_{\mathcal{F},p:p+q}$ from the LLM's tokenizer $T_\mathcal{F}$, with $j, l \geq 1$.[2] The relevance score for each LLM token in $y_{\mathcal{F},p:p+q}$ is computed as the average relevance score over the aligned FCM span $y_{\phi,l:l+m}$: $W_{p:p+q} = \{\frac{\sum_{l \leq t < m}(u_t)}{|y_{\phi,l:l+m}|} \mid y_t \in y_{\mathcal{F},p:p+q}\}$ In the example of Figure 3, the weight for the LLM's token *McDowall* is computed as the average over the weights for the three aligned FCM tokens (*McD*, *ow*, *all*). Further, since both LLM tokens *K* and *ing* are aligned to the single token *King* of the FCM, both tokens are set to the same relevance weight according to our algorithm.

5 Evaluation

Datasets & Metrics. We conduct experiments on long-form QA datasets of the ALCE citation benchmark (Gao et al., 2023b), namely their version of ASQA (Stelmakh et al., 2022) and ELI5 (Fan et al., 2019). We further evaluate models on BIO (Min et al., 2023). For our domain transfer experiments, we further consider Hagrid (Kamalloo et al., 2023) as a source for training. Since Hagrid's answers are generated by GPT-4 without annotations for factual coverage, we do not use it for evaluation as a target. For training, ALCE contains $\mid \mathcal{D} \mid = 4$ cited gold instances. For Hagrid, we randomly sample 4 instances for CaLF. Details are in Appendix A.1.

We follow the official metrics and nomenclature of Gao et al. (2023b) to measure correctness (via EM Recall), fluency (via MAUVE (Pillutla et al., 2021)), and citation F_1 (via an NLI-trained T5-11B model (Honovich et al., 2022)). We further measure Rouge-L (Lin, 2004) and introduce a more strict variation of their correctness metric: *passage-grounded correctness* (Correct. in $\mathcal{P}$), which only considers information from responses which are supported by the retrieved passages P. Subsequently, this metric ignores factual content produced by an LLM's parametric memory if not explicitly derivable from the retrieved passages, isolating factual grounding from a model's parametric memory. We use FactScore (Min et al., 2023) to evaluate the biographies of BIO. Detailed metric

[1] The exploration of alternative normalization functions is left to future work.

[2] Note that we implement this mapping function tailored to the specific tokenizers used by our models, see Appendix A.2.

Method		ALCE-ASQA					ALCE-ELI5				
		Similarity Rouge-L	Fluency MAUVE	Correct. EM Rec.	**Correct. in P**	**Citation F_1**	Similarity Rouge-L	Fluency MAUVE	Correct. EM Rec.	**Correct. in P**	**Citation F_1**
ChatGPT (Gao et al., 2023b)		–	66.6	40.4	–	73.1	–	57.2	12.0	–	50.5
GPT-4 (Gao et al., 2023b)		–	67.1	41.3	–	71.9	–	38.4	14.2	–	46.9
AGREE (Ye et al., 2024)		–	–	40.9	–	75.1	–	–	–	–	–
Self-RAG 7B (Asai et al., 2024)		35.7	74.3	30.0	–	67.3	16.9	32.6	9.7	5.4	27.6
BP, T5-3B (Fierro et al., 2024)		–	–	33.8	–	77.8	–	–	5.2	–	60.9
Llamav2-7B-chat	In-context	$35.9_{0.3}$	$77.8_{3.1}$	$35.0_{0.6}$	$25.7_{0.6}$	$49.9_{1.0}$	$20.5_{0.2}$	$36.2_{2.5}$	$17.7_{0.6}$	$10.8_{0.6}$	$38.2_{0.6}$
	Few-shot FT	$34.9_{0.4}$	$69.2_{4.3}$	$32.0_{0.4}$	$22.3_{0.7}$	$55.0_{1.8}$	$21.3_{0.2}$	$58.2_{2.2}$	$17.8_{0.6}$	$11.2_{1.1}$	$48.7_{2.9}$
	Ours	$37.8_{0.4}$	$86.0_{3.7}$	$37.7_{0.6}$	$29.3_{0.4}$	$70.4_{2.5}$	$20.8_{1.0}$	$59.6_{11.5}$	$17.0_{0.3}$	$11.9_{0.2}$	$66.5_{5.9}$
Mistral-Instr. 7B	In-context	$36.7_{0.3}$	$85.5_{2.7}$	$34.4_{0.4}$	$27.8_{0.4}$	$22.3_{0.9}$	$21.6_{0.9}$	$43.8_{4.8}$	$19.1_{0.4}$	$11.1_{0.2}$	$19.5_{0.6}$
	Few-shot FT	$38.1_{0.2}$	$\mathbf{87.7}_{0.8}$	$36.1_{0.9}$	$29.4_{1.1}$	$66.7_{4.5}$	$20.5_{0.3}$	$48.0_{4.9}$	$15.5_{1.2}$	$10.0_{0.6}$	$49.9_{4.0}$
	Ours	$37.2_{1.2}$	$84.5_{7.5}$	$36.4_{1.6}$	$30.0_{1.1}$	$76.2_{1.9}$	$\mathbf{21.8}_{0.2}$	$58.2_{4.0}$	$19.5_{0.8}$	$13.1_{0.2}$	$66.0_{4.7}$
MistralOrca 7B	In-context	$38.7_{0.1}$	$54.7_{1.8}$	$40.2_{0.3}$	$31.9_{0.2}$	$55.6_{0.8}$	$20.9_{0.1}$	$29.3_{0.8}$	$\mathbf{20.8}_{0.4}$	$12.5_{0.5}$	$43.3_{0.5}$
	Few-shot FT	$38.4_{1.8}$	$78.6_{14.7}$	$38.4_{3.8}$	$29.9_{4.7}$	$62.6_{3.6}$	$19.4_{1.9}$	$60.5_{13.6}$	$17.3_{1.8}$	$10.9_{1.2}$	$57.7_{6.5}$
	Ours	$\mathbf{40.3}_{0.2}$	$84.0_{3.3}$	$\mathbf{41.7}_{1.2}$	$\mathbf{34.5}_{0.5}$	$\mathbf{81.5}_{2.5}$	$20.4_{1.5}$	$\mathbf{62.7}_{4.6}$	$18.4_{2.1}$	$\mathbf{13.1}_{0.7}$	$\mathbf{73.1}_{4.2}$

Table 1: Main results on ALCE using only 4 training samples as $\mathcal{D}$, measured across 3 random seeds. Fine-tuning via CaLF substantially improves citation quality (**Citation F_1**) and correct information grounded in passages (**Correct. in P**) over alternative training strategies and competitive baselines while maintaining factual coverage (**EM Recall**), fluency (**MAUVE**), and recall-oriented similarity (**ROUGE-L**) to gold responses without citations.

descriptions (e.g. citation recall and precision) and results are in Appendix A.3.

Experimental Setup. Following recommendations for weakly supervised learning (Zhu et al., 2023) and few-shot learning (Alex et al., 2021), we do not consider a validation set for hyperparameter-tuning, representing real-world scenarios more accurately. Due to computational constraints, we use LoRA (Hu et al., 2022) for parameter-efficient fine-tuning of CaLF and our fine-tuning baselines. We use Alignscore as our FCM ϕ in all experiments unless otherwise mentioned. A threshold is used to map Alignscore's output o into a binary prediction for $\mathcal{Q}$. Alignscore is substantially different from the FCM model used for evaluating citations, using a different architecture and training data (c.f. App. A.2). ASQA, Hagrid, and BIO use Wikipedia as their underlying knowledge base while ELI5 uses CommonCrawl. We use the same retrievers as Gao et al. (2023b) and Asai et al. (2024) to maintain comparability. Further implementation details are in Appendix A.2.

Baselines. We compare CaLF on identical instruction-tuned LLMs against both in-context prompting and few-shot fine-tuning (Few-shot FT) via Eq. 1, trained on $\mathcal{D}$. In our domain transfer experiments, the entire 335 training samples of Hagrid are used to train the fine-tuning baseline (FT). Furthermore, we consider state-of-the-art models, namely the most powerful in-context prompted baselines from Gao et al. (2023b), ChatGPT (`gpt-3.5-turbo-0301`) and GPT-4 (`gpt-4-0613; 8K context window`). Due to context length limitations, in-context prompting uses 2 randomly sampled instances. We further compare against the best results of AGREE (Ye et al., 2024) which use PaLM 2's text-bison-001. Finally, we evaluate against open-source models, including Self-RAG 7B (Asai et al., 2024), based on Llama2, and Blueprint (BP) (Fierro et al., 2024), based on T5-3B.

5.1 Main Results

Table 1 shows the main in-domain results, with mean and standard deviation computed over three seeds. CaLF improves citation F_1 across datasets and models by 34.1 and 15.5 points over both in-context learning (In-context) and baseline fine-tuning (Few-shot FT), respectively. Impressively, *all* tested LLMs with CaLF outperform Self-RAG, ChatGPT, and GPT-4 with an average improvement of 24.8, 10.5, and 12.9 citation points, respectively. Moreover, CaLF achieves high citation scores while the overall quality of the response remains high. We observe modest improvement in correctness but substantial gains in grounded correctness, indicating that CaLF is better at including verifiable facts in its responses than our baselines. This contrasts observations for other models: GPT-4 trades off improvements in correctness at the cost of citation quality, resulting in ChatGPT having overall higher citation scores than GPT-4. Notably, CaLF improves on correctness over GPT-4 while also producing higher-quality citations than ChatGPT, beating both models at their best-performing metric. Similarly to GPT, BP produces accurate citations but has low correctness, especially on ELI5.

Method Source→Target	Similarity RougeL	Fluency MAUVE	Correct. EM Rec.	Correct. in P	Citation F_1	Method Source→Target	Similarity RougeL	Fluency MAUVE	Correct. EM Rec.	Correct. in P	Citation F_1
Self-RAG 7B	35.7	74.3	30.0	–	67.3	Self-RAG 7B	16.9	32.6	9.7	5.4	27.6
Llama2-7B-chat											
Zero-Shot →ASQA	36.1	47.5	35.6	27.1	31.0	Zero-Shot →ELI5	20.0	26.5	15.3	9.7	24.4
Few-shot FT ELI5→ASQA	37.2	74.0	37.7	31.6	64.9	Few-shot FT ASQA→ELI5	17.1	20.8	11.4	6.8	32.9
Ours ELI5→ASQA	37.1	75.7	36.1	30.4	73.1	Ours ASQA→ELI5	**21.3**	35.0	18.2	11.0	36.5
MistralOrca-7B											
Zero-Shot →ASQA	39.0	78.9	39.5	31.6	5.7	Zero-Shot →ELI5	**21.3**	35.0	**22.2**	12.6	10.4
Few-shot FT ELI5→ASQA	39.7	**90.1**	38.5	31.4	71.7	Few-shot FT ASQA→ELI5	20.9	41.1	19.7	10.6	40.4
Few-shot FT Hagrid→ASQA	36.7	66.1	37.8	29.5	51.3	Few-shot FT Hagrid→ELI5	**21.3**	**58.1**	20.9	12.7	32.3
Ours ELI5→ASQA	**40.1**	86.6	**40.0**	**33.2**	79.5	Ours ASQA→ELI5	21.2	31.3	20.4	12.5	**57.3**
Ours Hagrid→ASQA	39.7	80.8	38.6	32.3	**80.0**	Ours Hagrid→ELI5	21.1	32.3	20.2	**12.9**	54.6

Table 2: Results for our zero-shot domain transfer setting, when trained on a source dataset ($\mathcal{D}$) and evaluated on a different target dataset without any in-context instances. CaLF's citation quality (**Citation F_1**) and passage-grounded correctness (**Correct. in** P) is superior to all baselines, without additional inference costs.

Moreover, training via CaLF also leads to an overall improvement in ROUGE-L and MAUVE across models over the fine-tuning baseline by an average of 1.0 and 5.5, respectively.

5.2 Domain Transfer

We run domain transfer experiments to evaluate the generalization of CaLF's citation production, by training an LLM with CaLF on a source dataset and measuring performance on a different target dataset. Table 2 shows our domain transfer results. In every source-target configuration and across instruction-tuned models, CaLF outperforms zero-shot in-context learning, Few-shot FT, FT, and Self-RAG in both citation quality and correctness. CaLF (using MistralOrca-7B) exhibits small variability regarding the training source $\mathcal{D}$, with a citation F_1 difference $\Delta_{*\rightarrow\text{ASQA}}$ of 0.5 and $\Delta_{*\rightarrow\text{ELI5}}$ of 2.7, respectively. While CaLF's citation quality is comparable in-domain versus in a transfer setting on ASQA ($-\Delta 3.9$, see Table 1), we observe larger differences on ELI5 ($-\Delta 11.9$), since for ELI5 the knowledge source and question scope greatly differ from the Wikipedia-based source datasets.

5.3 Factuality

We evaluate CaLF's factual precision using FactScore. Results are shown in Table 3 for state-of-the-art methods taken from Asai et al. (2024), our fine-tuning baseline, and CaLF with MistralOrca-7B. CaLFscores the highest with 83.4, indicating that the improved citation quality translates to higher factual accuracy. Interestingly, while citation recall can be considered a stricter measure than the FactScore, the former is much higher for CaLF. We postulate the difference is caused by retrieval inaccuracies. Subsequently, we adjust predictions to abstain from answering if none of the passages' titles match with the BIO entity.[3] As seen in Table 3, the FactScore improves to 86.1 and 88.9 for our fine-tuning baseline and CaLF, respectively. While these results are promising they also highlight the importance of accurate retrieval and high-quality knowledge bases so that citation production can translate into improved factuality, an observation also made in Menick et al. (2022); Kryscinski et al. (2020).

Method	FS	Citation Recall
ChatGPT w/o retrieval	71.8	–
Llamav2-13B-chat	79.9	–
Self-RAG 7B	81.2	–
Self-RAG 13B	80.2	–
Few-shot FT	78.7	69.3
Few-shot FT + Retrieval Filtering	86.1	69.3
Ours	83.4	92.7
Ours + Retrieval Filtering	**88.9**	**92.7**

Table 3: FactScore (FS) evaluation. MistralOrca7B is used with *Few-shot FT* and *Ours*. Other results are taken from Asai et al. (2024). + *Retrieval Filtering*: the model abstains when no passage title matches the entity.

6 Discussion

Ablation. Table 4 shows an ablation for the two mechanisms introduced in CaLF, the generation of weakly-supervised training data (WS) and the focused learning loss ($\mathcal{L}_{FL}$). By adding WS, we see an improvement of 18.3 and 16.7 citation recall and precision points on ASQA, respectively. By

[3] Abstaining is explicitly incorporated into FactScore.

adding $\mathcal{L}_{FL}$, we further improve on top of WS by 6.3 and 8.2 recall and precision points, respectively. On ELI5, we observe similar improvements.

Method	Correctness in P	Citation Recall	Citation Precision
ASQA			
LLM	25.8	57.5	55.2
LLM + WS	29.0 (+3.2)	75.8 (+18.3)	71.9 (+16.7)
LLM + WS + $\mathcal{L}_{FL}$	**33.8** (+8.0)	**84.1** (+24.6)	**83.8** (+28.9)
ELI5			
LLM	11.0	57.3	51.0
LLM + WS	11.5 (+0.5)	61.5 (+4.2)	57.2 (+6.2)
LLM + WS + $\mathcal{L}_{FL}$	**12.5** (+1.5)	**72.1** (+14.8)	**66.6** (+15.6)

Table 4: Ablation for CaLF (MistralOrca-7B). WS: weakly-supervised training, $\mathcal{L}_{FL}$: Focused learning.

Iterative Training. Performance of CaLF across iterations is shown in Figure 4 on ASQA. We see a majority of citation performance improvements within the first three iterations after which citation F_1 stabilizes. We further observe consistent improvements to MAUVE, ROUGE-L, and grounded correctness across iterations. Importantly, we do not observe erratic or unstable performance, indicating the robustness of our iterative training procedure. The proportion of filtered examples $\hat{\mathcal{D}}_k$ over $\tilde{Y}_k$ as our dynamic stopping criterion matches the citation performance well, improving efficiency by early stopping once saturated (here iteration 4).

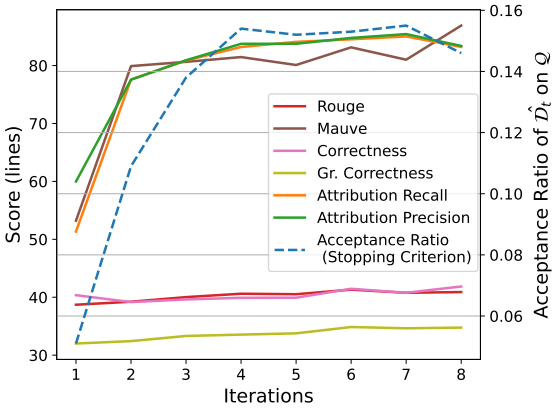

Figure 4: Evaluation metrics and CaLF's dynamic stopping criterion over the number of iterations on ASQA.

Selection of FCM. We further investigate the extent to which the quality of an FCM translates to improved citation production for CaLF. To this end, we replace the FCM with: (i) a DeBERTav3 model adjusted for factual consistency (Steen et al., 2023), (ii) the current T5-based state-of-the-art on the TRUE benchmark (Gekhman et al., 2023), (iii) the FCM Honovich et al. (2022) used for citation evaluation. We constrain $\hat{\mathcal{D}}_t$ to 32 samples and run without $\mathcal{L}_{\text{FL}}$ since for decoders the computation of Shapley values over their vocabulary (prediction) is more involved. Table 5 shows results on the TRUE benchmark (summarization subset) and the citation F_1 on ASQA with CaLF. We generally observe that better scores on the TRUE benchmark translate to higher citation scores when incorporated into CaLF. The only exception is the FCM Honovich et al. (2022), which is used in both training and evaluation, scoring disproportionately high in terms of citation F_1, most likely due to model biases leaking through evaluation. While Gekhman et al. (2023) performs the best on TRUE, it is much more expensive to run at scale for CaLF than AlignScore.

Method	TRUE AUC	Citation F_1
Steen et al. (2023)	79.5	72.5
Honovich et al. (2022)†	82.7	81.9
AlignScore (Zha et al., 2023)	84.8	74.0
Gekhman et al. (2023)	87.8	77.7

Table 5: Measuring the extent to which the quality of an FCM (TRUE benchmark) translates to CaLF by comparing the quality of an FCM (TRUE benchmark) and CaLF. †Citation evaluation model.

Adversarial Baselines. Long-form QA is notoriously difficult to automatically evaluate (Xu et al., 2023; Krishna et al., 2021). Thus, to assess whether the automatic evaluation metrics of Gao et al. (2023b) can be exploited, we designed several adversarial baselines, such as copying retrieved passages as the answer or citing all passages for every generated sentence. Results are shown in Table 6. Every baseline that attempts to trick one metric performs poorly on another. For instance, copying passages (Copy) results in very poor MAUVE while citing all passages (Ours w/ cite all) results in very low citation precision. A more detailed discussion is shown in Appendix A.3.

Human Evaluation. We further conduct a human evaluation to compare the quality of generated responses produced by CaLF with Few-shot FT (using MistalOrca-7B). Human subjects are tasked to judge the models' generations regarding citation quality, informativeness, coherence, and fluency, following human evaluation in Gao et al. (2023b) and recommendations in Zhong et al. (2022). The latter three metrics are measured using a five-point Likert scale, with five being the highest score. We

Method	Similarity Rouge-L	Fluency MAUVE	Correct. EM Rec.	Correct. in P	Citation Recall	Citation Precision	Avg length
Ours	40.5	80.1	39.9	33.8	84.1	83.8	71.2
Copy (No length limit)	25.3	11.8	50.6	49.5	98.2	98.7	628.4
Copy (Truncated - 1st paragraph)	34.0	16.8	36.0	34.6	97.9	98.4	210.7
Copy (Truncated - 100 tokens)	33.3	16.2	22.8	20.9	96.1	96.3	100.0
Ours w/ all set to [1]	40.5	80.1	39.9	33.8	49.0	49.5	71.2
Ours w/ cite all	40.5	80.1	39.9	33.8	75.6	36.4	71.2
Ours w/o EOS token	33.0	20.2	46.9	39.7	71.1	81.7	256.0

Table 6: Evaluation of adversarial baselines on ASQA, using MistralOrca-7B. Results are shown for seed 42. *Copy*: Use retrieved passages directly as a response. *w/ all set to [1]*: Replacement of all citations in the generated response with citations to the first passage. *w/ cite all*: Replacement of citations in the generated response with citations to all passages. *w/o EOS token*: Remove EOS token from training. The average gold answer length is 113 tokens.

randomly sample 30 instances for each model from ELI5, ASQA, and BIO, resulting in 180 instances. Results are shown in Table 7. CaLF achieves substantially higher citation quality, fluency, and answer coherence ratings than across datasets with an F_1 of 91.5 versus 72.7 for the baseline. While average informativeness ratings are comparable between CaLF and the baseline , on ELI5 particularly we observe low informativeness with CaLF. This is caused by frequent retrieval errors since ELI5 uses a weak retriever (BM25) to efficiently traverse over its large and noisy knowledge base. Instead of using its parametric memory when passages contain little relevant information, we observe that CaLF still produces accurately grounded responses, however, at the cost of less informative content.

Method		Citation F_1 of 100 ↑	Informativeness Likert Scale 1 to 5 ↑	Coherence Likert Scale 1 to 5 ↑	Fluency Likert Scale 1 to 5 ↑
Few-shot FT	ASQA	65.2	4.07	4.07	4.57
	ELI5	68.7	4.13	4.07	4.5
	BIO	84.3	4.6	4.57	4.67
	Avg	72.7	4.27	4.23	4.58
Ours	ASQA	93.7	4.67	4.67	4.93
	ELI5	82.7	3.3	3.93	4.17
	BIO	97.3	4.87	4.77	4.83
	Avg	**91.5**	**4.28**	**4.46**	**4.64**

Table 7: Human evaluation study. Informativeness, coherence, and fluency are judged using a five-point Likert scale. Results confirm automatic evaluation, putting CaLF ahead in terms of citation quality while maintaining high response quality.

Efficiency At inference time, CaLF matches the efficiency of the few-shot FT baseline, contrasting previous methods that cause significant overhead, including in-context learning (increased input length), Self-RAG (tree-decoding with critique tokens), and post-hoc editing. The training complexity of CaLF can be described as $\mathcal{O}(K \times |\mathcal{X})|)$, with K and $\mathcal{X}$ being the number of iterations and the collection size of questions we generate weakly-supervised data of, respectively. The generation of diverse answer candidates is computationally the most involved while the filtering of answer candidates and the computation of Shapley Values is efficient due to the small size of the FCM we use (Alignscore, 355M parameters). To quantify the computational cost, we measured training times using a single A100 40GB GPU. We measure a training time of 13h51min for CaLF and 1h2min for the few-shot FT baseline. While there's a notable disparity in training time, it's essential to note that achieving comparable performance to CaLF via regular fine-tuning would necessitate training on significantly more data, resulting in additional training and, importantly, data annotation cost.

7 Conclusion

This paper presented CaLF, a fine-tuning method for LLMs to produce accurate citations alongside generated text. It focuses on using FCMs as a training signal by filtering candidate answers with citations and by re-weighting the LLM's objective function according to the tokens' factual importance. CaLF outperforms all baselines in terms of citation quality and passage-grounded correctness while ensuring that the overall quality of responses remains high, measured via both automated and human evaluation. In a domain transfer setting we further validate the generalizability of CaLF. Moreover, we discuss the benefits of accurate citation production for improving factuality and highlight the importance of each component of CaLF through a systematic ablation. Future work looks at incorporating a learned mechanism into CaLFto abstain from answering if none of the retrieved passages are considered relevant to the question.

Limitations

The assumption that every generated sentence requires a citation is an oversimplification we adopt from Gao et al. (2023b). However, real-world dialogue agents commonly introduce their response with non-factual phrases or sentences (e.g. *Of course I can help!*). A potential solution is to introduce an extrapolatory label as described in Yue et al. (2023). Furthermore, CaLF is not entirely model-agnostic out-of-the-box, since the tokenization alignment algorithm was designed with the idiosyncrasies of the particular LLMs and FCMs in mind. Moreover, inaccuracies or biases of the factual consistency models could negatively affect the citation quality of CaLF (c.f. Sec. 6). A common strategy to mitigate biases involves an alignment stage, where models are explicitly optimized on based on preference data. It would be interesting to explore alignment in the context of CaLF applied to highly domain-specific questions, such as those compiled in the recent dataset of Malaviya et al. (2024).

Finally, our paper focuses exclusively on improving generation, yet as highlighted in Section 6, high-quality retrieval systems are vital for citations to be effective. When retrieved passages contain non-factual information or lack useful content, produced answers might not be informative. While the reliance of retrieved passage is answer production poses a limitation, language model's intrinsic memory for content generation forfeits the transparency and verifiability benefits inherent in citation production. Balancing the use of citation production and intrinsic memory while preserving the advantages of both remains an ongoing challenge. Future directions might consider a joint fine-tuning procedure of generator and retriever and even capturing their rich interactions, such as *abstaining* from answering questions when retrieved passages are irrelevant. The action of abstaining to answer could serve as a signal to the retrieval system to seek alternative sources of information.

Ethics Statement

Our paper improves the ability of LLMs to produce accurate citations to sources alongside their generated answers to improve the verifiability of LLMs' responses. As noted in Section 5.3, we caution against equating improved citation quality with improved factuality since even correctly cited passages can be factually incorrect or misleading without additional context about the passage and its source. Citations can therefore invoke a false sense of trust and amplify observations made in Si et al. (2023). It remains important for users to stay critical and reflect on generated responses. By improving citation accuracy our paper simplifies the verification process but does not eliminate it. Furthermore, Huang and Chang (2024) argue that the incorporation of citations can help to address intellectual property and associated ethical issues in the deployment of LLMs due to the increased verifiability of responses made. Finally, we recognize a potential risk of dual-use by adversarial actors: similarly to how humans cherry-pick data to support a strong bias or prior, we cannot guarantee that an LLM will not exhibit similar behaviour when manipulated with passages that contain misinformation or miss relevant context.

Acknowledgements

The authors would like to thank the anonymous reviewers for their time and detailed feedback on our paper. We further thank Tony Hu for being part of the human evaluation.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. 2021. RAFT: A real-world few-shot text classification benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev,

Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Yifan Deng, Xingsheng Zhang, Heyan Huang, and Yue Hu. 2023. Towards faithful dialogues via focus learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4566, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations.

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 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

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

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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. 2024. Citation: A key to building responsible and accountable large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4940–4957, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

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

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

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Lloyd S Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, NLI models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924, Toronto, Canada. 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.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

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*.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Xi Ye, Ruoxi Sun, Sercan Ö Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Datasets

The four datasets considered for evaluation cover different question-answering tasks and domains: ASQA is a disambiguation task built on top of AmbigQA (Min et al., 2020), ELI5 contains real-world highly open-ended questions and answers from an online forum (Reddit), Hagrid contains entity-specific questions, and BIO is a biography generation task containing simple person-related questions. Hagrid's training set as provided by Kamalloo et al. (2023) consists of 1,922 GPT-4 generated answers, out of which 335 instances are human-labeled as both attributable and informative. Only these 335 instances are used as training data. For Hagrid, we randomly sample 4 instances for CaLF and use all its 335 training samples for our fine-tuning baseline. We use each datasets' original training set to sample questions $\mathcal{X}$ and atomic facts A. Since Hagrid does not have any annotations for factual coverage, we compute the quality assurance function $\mathcal{Q}$ only over the Citation conditions. BIO consists of two evaluation sets. As recommended by the authors and following Self-RAG, we use the second, and more difficult, evaluation set in our experiments. The test set of ALCE-ASQA and ALCE-ELI5 consists of 1000 randomly sampled instances from their respective development set. For more details on the ALCE benchmark, we refer to Gao et al. (2023b).

A.2 Experimental Setup

Token alignment algorithm We design the token alignment algorithm around the idiosyncrasies of the LLMs' and FCM's tokenizers. Llama2, Mistral-Instruct, and MistralOrca use the Llama tokenizer[4], a BPE model (Sennrich et al., 2016) based on sentencepiece[5]. Similarly, the FCM, Alignscore, uses a RoBERTa tokenizer which is derived from the GPT-2 tokenizer, also using BPE. Both Llama and RoBERTa tokenizers treat spaces as parts of the tokens. In contrast to RoBERTa tokenizer, the Llama tokenizer indicates each space via an underscore ("_"). We exclude special tokens of either tokenizer (e.g. "<s>" or "<0x0A>") from the alignment procedure.

Given a sentence, tokenized both by FCM and LLM, we we first strip the FCM tokens (i.e. remove spaces) and remove underscores for tokens of the LLM, essentially removing all spaces to unify their representations. The algorithm then checks whether the current tokens are equal or subsets of one another, with and without considering the memory of tokens not yet aligned but iterated over by the pointers. For instance, consider the FCM tokens: 'Maw', 'syn', 'ram', and LLM tokens for the same word: is _M', 'aws', 'yn', 'ram'. For the first two FCM tokens, none of the LLM tokens can be directly aligned to it. Our algorithm ensures to find the *smallest sequence* of tokens in both the LLM and FCM that can be aligned to each other (here the spans are 'Mawsyn' and 'ram') and assigns the relevance score as described in Section 4.2. We find the smallest sequence by keeping a list of tokens for each FCM and LLM that could not yet be aligned and increment the respective pointers according to which sequence needs to continue (e.g. Checking that [M] + [was] continues [Maw] potentially, requires incrementing the FCM pointer, since an 's' is missing in [Maw]. This algorithm is efficient and runs in linear time.

Generating Cited Answers for Training. While datasets such as ELI5 have over 250K training instances we could use to generate weakly-supervised training data, we constrained the size of $\mathcal{X}$ to 1000 samples, since it is computationally infeasible to run our weakly-supervised data generation procedure on all training samples of the dataset. The threshold Θ is determined dynamically. Starting with a value of 0.9, if the number of samples in $\hat{D}$ is below 3, we consider the threshold too high for the given LLM and reduce it by 0.1 until the requirement is met.

Passage Retrieval. For the retriever $\mathcal{R}$ we use GTR (Ni et al., 2021) for ASQA and Hagrid (Wikipedia), BM25 for ELI5 (CommonCrawl[6]), and Contriever-MS MARCO (Izacard et al., 2022) for BIO (Wikipedia), to maintain comparability with Gao et al. (2023b) and Asai et al. (2024). We use $|P| = 3$ passages throughout due to context length limitations. We use the same indices as provided by ALCE, subsequently, each passage has a length of 100 tokens.

Answer Generation. We use the same instructions/prompts across all models, specifically the

[4] https://huggingface.co/docs/transformers/main/en/model_doc/llama#transformers.LlamaTokenizer

[5] https://github.com/google/sentencepiece?tab=readme-ov-file

[6] http://commoncrawl.org

default prompts from ALCE. In contrast to the code of ALCE[7], we used the chat templates across all models and baselines[8] to render the inputs, since we observed crucial tokens missing during the fine-tuning procedure otherwise. The only exception is ELI5, where we observed that chat templates resulted in worse performance and used ALCE's strategy plus relevant formatting tokens instead across all models and baselines. For inference, we use MAP with a beam size of 1, instead of sampling with temperature scaling as done in ALCE, assuming that this results in more factual outputs.[9]

Factual Consistency Model Our choice of using AlignScore as our FCM ϕ was further motivated by its dissimilarity to the citation evaluation model (Honovich et al., 2022) (in contrast to e.g. TrueTeacher (Gekhman et al., 2023) which in principle performed better but is more similar to the evaluation model). While AlignScore is an encoder-based RoBERTa model, TRUE (Honovich et al., 2022) is an encoder-decoder T5-11B model. Moreover, their training data is different. AlignScore was trained on 15 datasets of various tasks, including natural language inference, summarization, information retrieval, and paraphrasing while TRUE has only seen 6 natural language inference-related datasets.

Training & Hyperparameters We set the learning rate to 3^{-4} and train for a total of 100 steps across all models and experiments. The maximum generation length is set to 256 tokens, however, most generations stay far below this limit (see Appendix A.3. We use adamw (Loshchilov and Hutter, 2019) as the optimizer. We use a batch size of 1 during training with gradient accumulation, resulting in an effective batch size of 4. For LoRA, we use a rank $r = 4$ and apply it to all parts of the attention mechanism. For fine-tuning, we exclude tokens of the prompts from the loss computation that are not part of the gold answer, so we are not fine-tuning the instructions, only the answers that follow after the instruction. The iterative training process stops after at most 8 iterations due to computational constraints. We further down-weight the loss contribution of the EOS token to 0.02 since models tend to otherwise produce very short single-sentence responses.

Implementation Details We use the Huggingface checkpoints for LLama2-7B[10], MistralOrca-7B[11], and Mistral-Instruct-7B,[12] as well as for all FCMs we considered in our experiments, namely AlignScore (RoBERTa-large, 355M parameters)[13], TRUE (T5, 11B parameters)[14], (Steen et al., 2023)[15] (DeBERTaV3, 304M parameters), and TrueTeacher (T5, 11B parameters)[16]. The Mistral models are licensed under Apache2.0, AlignScore and (Steen et al., 2023) are licensed under MIT, TrueTeacher is licensed under cc-by-nc-4.0, and Llama2 is licensed under the llama license[17]. Subsequently, our research is consistent with the licenses' intended use. The models are intended to be used for English. We use the ALCE data and prompts from their repository[18], and Huggingface's Datasets for Hagrid[19]. For running the experiments, we used a combination of A100 40GB and A10G with 23GB GPUs. We use the Python package rouge-score[20] for computing the ROUGE, and the package mauve-text[21] for computing the MAUVE score. We adopt the code of ALCE for the citation and correctness metrics and define passage-grounded correctness ourselves. We use NLTK (Bird and Loper, 2004) for some pre-and post-processing steps.

Baselines. Results from state-of-the-art methods are taken from their respective papers. The only exception is Self-RAG on ELI5, which we have run by ourselves using the authors' repository and

[7] https://github.com/princeton-nlp/ALCE

[8] https://huggingface.co/docs/transformers/main/en/chat_templating

[9] Note that we do not use constrained decoding for citation generation despite its apparent suitability here. Yet, the position of citation markers within a sentence can be rather flexible, which we consider a desired property for naturally produced text.

[10] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[11] https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca

[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[13] https://huggingface.co/yzha/AlignScore, and their repository https://github.com/yuh-zha/AlignScore

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

[15] https://huggingface.co/juliussteen/DeBERTa-v3-FaithAug

[16] https://huggingface.co/google/t5_11b_trueteacher_and_anli

[17] https://github.com/facebookresearch/llama/blob/main/LICENSE

[18] https://github.com/princeton-nlp/ALCE

[19] https://huggingface.co/datasets/miracl/hagrid

[20] https://pypi.org/project/rouge-score/

[21] https://pypi.org/project/mauve-text/

their models[22]. At the point of submission, their repository produces results which are worse than reported in their paper, as explained by the authors due to a bug[23], which potentially impacts scores reported of Self-RAG on ELI5. While the authors have stated their intentions to fix this issue, the problem was not resolved as of our submission date. We will update their scores once a fix is available.

A.3 Evaluation

Detailed descriptions for each evaluation metric are shown in Table 8. Results shown in Table 2 are computed with default random seed 42.

Ablation Table 9 Table 10 show the ablation results for CaLFon Mistral-Instruct 7B and Llamav2-7B-chat, respectively. The results largely align with observations made for MistralOrca-7B in table 4. We observe that passage-grounded correctness either slightly decreases or remains comparable to the baseline when using the weakly-supervised training without our focused learning objective $\mathcal{L}_{\text{FL}}$. Once the objective is added,vcorrectness improves consistently across datasets and models over the baseline.

Adversarial Baselines The automated metrics proposed by Gao et al. (2023b) do not explicitly control for the generation of irrelevant information (i.e. factual precision, such as FactScore)[24]. Subsequently, their metrics favour longer responses for coverage-based measures (i.e. correctness), as also pointed out by Asai et al. (2024). This raises the question of whether the automated metrics deployed can be tricked with trivial responses or certain response patterns.

Table 6 shows results across metrics for several such baselines on ASQA. Our approach has an average response length of 71.2 tokens, comparable to typical fine-tuning with 61.0 tokens, both being substantially shorter than the dataset's average gold answer length with 113 tokens. In contrast, the retrieved passages themselves are 521.8 tokens long. Considering these as the answer themselves, we indeed achieve a higher correctness score (50.6) while maintaining perfect citation, however, we observe a substantial decrease in both ROUGE-L and MAUVE scores. This is intuitive since the retrieved passages are very dissimilar in style from the gold answers. Moreover, MAUVE has an explicit length bias (Pillutla et al., 2021). When explicitly biasing our CaLF to generate long sequences by removing the EOS token during fine-tuning and setting the generation limit to 256 tokens, we also observe a substantial increase in correctness while maintaining much of the attribution performance. Yet, again we observe a substantial decrease in ROUGE-L and MAUVE, highlighting the importance of maintaining all performance metrics high while optimizing attribution, as achieved by CaLF. Finally, we consider replacing the citations made in generated responses from our model with: (i) citations to exclusively the first passage, (ii) citations to all passages. As seen in the table, citation scores are substantially worse than our approach. While in principle (ii) should present an upper bound on attribution recall, we observe lower scores than our model even here. This can be explained by inaccuracies in the evaluation model, being biased towards information at the beginning of the premise (scoring much worse for citations to the last passage).

A.4 Human Evaluation

Human subjects are tasked to judge the models' generations regarding: (i) citation recall: judgement whether a generated sentence is fully supported by citations, (ii) citation precision: whether a citation partially or fully supports a sentence, (iii) informativeness: whether the generation helps to answer the question, (iv) coherence: whether generated sentences are semantically and syntactically well-connected, (v) fluency: whether all sentences are grammatically correct and well-readable. While (i), (ii), and (iii) are adopted from the human evaluation in Gao et al. (2023b), we further measure (iv) and (v) as recommended in Zhong et al. (2022). We randomly sample 30 instances for each model for each dataset (ELI5, ASQA, BIO), resulting in 180 instances which are judged via the above criteria. The instances were judged by four subjects, distributed equally (45 samples per subject). Each subject was provided with detailed annotation guidelines, providing examples for each possible annotation option with explanations. The subjects were partially authors and partially volunteers. All subjects were informed how the data would be used and provided consent. All subjects are male under 65 years of age and either from the USA or the UK.

[22] https://github.com/AkariAsai/self-rag

[23] https://github.com/AkariAsai/self-rag/issues/4

[24] The exclusion of such metric in ALCE is likely due to gold answers not being designed to be factually comprehensive.

Name	Description	Computation
Gold Answer-based Metrics		
Fluency (MAUVE)	Measures two types of errors: (i) model produces degenerate text (outside of human distribution), (ii) model does not yield diverse text (does not cover human distribution)	$R_\lambda = \lambda P + (1-\lambda)Q$, summarizing KL divergence of KL$(P\|R_\lambda)$ and KL$(Q\|R_\lambda$, for $\lambda \in (0,1)$, computed via Monte-Carlo estimator (LM embeddings + quantization via k-means).
Similarity (ROUGE-L)	Measures longest matching sequence of words. Bad approximator for factuality (see e.g. here)	Computes F_1 for longest common subsequence (LCS). Recall being ratio over reference. Precision ratio over answer.
Specialized Metrics		
Correctness	Measures whether atomic units of information from gold answers appear in the generated answer.	ASQA: Measures exact match of short answers in generated answers. ELI5: Measures whether atomic statements are consistent with generated answer.
Passage-grounded Correctness	Measures whether atomic units of information from gold answers appear in the generated answers *and* whether this information is supported by the retrieved passages. Eliminates to score responses that are correct but not grounded in retrieved passages.	ASQA: Measures exact match of short answers in generated answers that can be attributed to retrieved passages. ELI5: Measures whether atomic statements are consistent with generated answer that can be attributed to retrieved passages .
Citation Recall	Measures the ratio of sentences in generated answers that are consistent with their cited sources/passages.	Recall is 1 iff there exists at least one citation and sentence is consistent with citations considered jointly: $\phi_{eval}(\oplus(C_i), s_i) = 1$.
Citation Precision	Measures the ratio of citations in generated answer that are not irrelevant.	Citation considered irrelevant iff: i) the citation itself does not support the attributed sentence: $\phi_{eval}(c_{i,j}, s_i) = 0$ ii) removing the citation does not affect rest of citations to attribute the sentence: $\phi_{eval}(\oplus(C_i)\backslash\{c_{i,j}\}, s_i) = 1$.

Table 8: Overview of the LFQA evaluation metrics. Correctness, Fluency, and Citation scores are taken from the ALCE (Gao et al., 2023b). Passage-grounded Correctness is a metric we propose to measure information coverage for citable statements (i.e. excluding hallucinated correct information). ϕ_{eval} is the evaluation FCM, namely TRUE (Honovich et al., 2022).

Method	Correctness in P	Citation Recall	Citation Precision
ASQA			
LLM	28.3	60.9	63.5
LLM + WS	24.1 (-4.2)	72.3 (+11.4)	73.9 (+10.4)
LLM + WS + $\mathcal{L}_{FL}$	**29.6** (+1.3)	**79.2** (+19.7)	**80.2** (+16.7)
ELI5			
LLM	9.1	55.2	40.1
LLM + WS	10.9 (+1.8)	60.9 (+5.4)	59.6 (+19.5)
LLM + WS + $\mathcal{L}_{FL}$	**12.5** (+3.4)	**70.0** (+14.8)	**67.0** (+26.9)

Table 9: Ablation for CaLF (Mistral-Instruct-7B). WS: weakly-supervised training, $\mathcal{L}_{FL}$: Focused learning.

Method	Correctness in P	Citation Recall	Citation Precision
ASQA			
LLM	23.4	58.7	55.3
LLM + WS	23.1 (-0.3)	69.6 (+10.9)	66.3 (+11.0)
LLM + WS + $\mathcal{L}_{FL}$	**30.7** (+7.3)	**76.0** (+17.3)	**72.5** (+17.2)
ELI5			
LLM	11.3	53.2	46.6
LLM + WS	9.1 (-2.1)	67.1 (+13.9)	66.4 (+19.8)
LLM + WS + $\mathcal{L}_{FL}$	**11.7** (+0.4)	**71.2** (+18.0)	**63.2** (+16.6)

Table 10: Ablation for CaLF (Llama2-7B-chat). WS: weakly-supervised training, $\mathcal{L}_{FL}$: Focused learning.

A.5 Qualitative Examples

To qualitatively compare the produced answers by CaLF and our few-shot FT baseline we show a randomly selected output for each evaluation dataset. Results are shown in Figure 5, Figure 6, and Figure 7 for ASQA, ELI5, and BIO, respectively.

Method	Rouge-L	Fluency	Correctness	Grounded Correct.	Citation Recall	Citaion Precision	Avg length
ASQA							
ChatGPT (Gao et al., 2023b)	–	66.6	**40.4**	–	73.6	72.5	–
Vicuna-13B (Gao et al., 2023b)	–	82.6	31.9	–	51.1	50.1	–
Self-RAG 7B (Asai et al., 2024)	35.7	74.3	30.0	–	66.9	67.8	–
In-context (Llamav2-7B-chat)	35.9	84.1	34.5	25.2	50.4	50.0	93.3
Few-Shot FT (Llamav2-7B-chat)	34.7	71.3	33.1	23.4	58.7	55.3	52.5
In-context (Mistral-Instruct 7B)	36.4	86.6	34.1	28.0	21.7	23.6	75.9
Few-Shot FT (Mistral-Instruct 7B)	37.6	82.5	35.3	28.3	60.9	63.5	105.7
In-context (MistralOrca 7B)	38.8	53.9	40.1	32.4	52.5	61.0	64.5
Few-Shot FT (MistralOrca 7B)	37.5	81.9	37.3	25.8	57.5	55.2	61.0
Ours (Llamav2-7B-chat)	37.9	84.3	37.8	29.5	72.8	72.3	86.9
Ours (Mistral-Instruct 7B)	37.7	**87.5**	35.2	29.6	79.2	80.2	79.5
Ours (MistralOrca 7B)	**40.5**	80.1	39.9	**33.8**	**84.1**	**83.8**	71.2
ELI5							
ChatGPT (Gao et al., 2023b)	–	57.2	12.0	–	51.1	50.0	–
Vicuna-13B (Gao et al., 2023b)	–	58.2	10.0	–	15.6	19.6	–
Self-RAG 7B (Asai et al., 2024)	16.9	32.6	9.7	5.4	23.3	33.9	–
In-context (Llamav2-7B-chat)	19.7	37.7	14.1	8.6	39.9	27.6	110.7
Few-Shot FT (Llamav2-7B-chat)	21.3	54.9	17.8	11.3	53.2	46.6	138.4
In-context (Mistral-Instruct 7B)	20.5	62.3	17.4	11.7	43.8	44.9	111.2
Few-Shot FT (Mistral-Instruct 7B)	19.5	37.8	13.8	9.1	55.2	40.1	93.0
In-context (MistralOrca 7B)	20.8	27.7	**20.5**	12.4	45.4	41.8	94.8
Few-Shot FT (MistralOrca 7B)	20.5	44.9	18.4	11.0	57.3	51.0	100.8
Ours (Llamav2-7B-chat)	21.3	**69.5**	17.2	11.7	71.2	63.2	141.2
Ours (Mistral-Instruct 7B)	**21.8**	53.5	18.9	**12.5**	70.0	**67.0**	143.8
Ours (MistralOrca 7B)	20.7	68.3	18.6	**12.5**	**72.1**	66.6	108.3

Table 11: Main in-domain results on ASQA, ELI5 using CaLF for fine-tuning various instruction-tuned LLMs, using only 4 initial samples $\mathcal{D}$. Results are shown for default random seed 42.

Method	Source	Rouge	Fluency	Correct.	Gr. Correct.	Attr. Recall	Attr. Precision	Length
Target Dataset: ASQA								
Self-RAG 7B	Critique tokens	35.7	74.3	30.0	–	66.9	67.8	–
Zero-Shot (Llama2-Chat-7B)	–	36.1	47.5	35.6	27.1	24.3	42.8	126.9
Zero-Shot (MistralOrca)	–	39.0	78.9	39.5	31.6	5.3	6.1	63.6
Few-shot FT (Llama2-Chat-7B)	ELI5	37.2	74.0	37.7	31.6	67.9	62.2	150.6
Few-shot FT (MistralOrca)	ELI5	39.7	**90.1**	38.5	31.4	73.8	69.7	94.1
FT (MistralOrca)	Hagrid	36.7	66.1	37.8	29.5	50.8	51.8	57.6
Ours (Llama2-Chat-7B)	ELI5	37.1	75.7	36.1	30.4	77.8	69.0	141.6
Ours (MistralOrca)	ELI5	**40.1**	86.6	**40.0**	**33.2**	**80.4**	78.5	101.0
Ours (MistralOrca)	Hagrid	39.7	80.8	38.6	32.3	78.9	**81.1**	64.9
Target Dataset: ELI5								
Self-RAG 7B	Instr. tuning w/ critique tokens	16.9	32.6	9.7	5.4	23.3	33.9	–
Zero-Shot (Llama2-Chat-7B)	–	20.0	26.5	15.3	9.7	17.8	39.0	111.1
Zero-Shot (MistralOrca)	–	**21.3**	35.0	**22.2**	12.6	8.1	14.4	120.5
Few-shot FT (Llama2-Chat-7B)	ASQA	17.1	20.8	11.4	6.8	33.3	32.6	70.4
Few-shot FT (MistralOrca)	ASQA	20.9	41.1	19.7	10.6	39.8	41.0	92.1
FT (MistralOrca)	Hagrid	**21.3**	**58.1**	20.9	12.7	28.6	37.0	112.4
Ours (Llama2-Chat-7B)	ASQA	**21.3**	35.0	18.2	11.0	38.6	34.7	137.7
Ours (MistralOrca)	ASQA	21.2	31.3	20.4	12.5	**57.4**	**57.2**	97.2
Ours (MistralOrca)	Hagrid	21.1	32.3	20.2	**12.9**	54.0	55.3	101.7

Table 12: Results for our zero-shot domain transfer setting, when trained on a source dataset and evaluated on a different target dataset without any in-context instances. Results are shown for default random seed 42.

Question: What is the genre of holes by louis sachar?
Retrieved Passages:
[1] Title: Louis Sachar Text: Louis Sachar Louis Sachar (; born March 20, 1954) is an American young-adult mystery-comedy author. He is best known for the "Wayside School" series and the award-winner, "Holes". "Holes" won the 1998 U.S. National Book Award for Young People's Literature and the 1999 Newbery Medal for the year's "most distinguished contribution to American literature for children". In 2013, it was ranked sixth among all children's novels in a survey published by "School Library Journal". After graduating from high school, Sachar attended Antioch College for a semester before transferring to University of California, Berkeley, during which time he began helping

[2] Title: Holes (novel) Text: American teenager with a criminal history. A new friendship with Ginny, who has cerebral palsy, a reunion with former friend X-Ray, a ticket-scalping scheme, a beautiful pop singer, and a frame-up all test Armpit's resolve to "Just take small steps and keep moving forward". Holes (novel) Holes is a 1998 young adult mystery comedy novel written by Louis Sachar and first published by Farrar, Straus and Giroux. The book centers around an unlucky teenage boy named Stanley Yelnats, who is sent to Camp Green Lake, a juvenile corrections facility in a desert in Texas, after being falsely accused of theft.

[3] Title: Holes (novel) Text: Holes (novel) Holes is a 1998 young adult mystery comedy novel written by Louis Sachar and first published by Farrar, Straus and Giroux. The book centers around an unlucky teenage boy named Stanley Yelnats, who is sent to Camp Green Lake, a juvenile corrections facility in a desert in Texas, after being falsely accused of theft. The plot explores the history of the area and how the actions of several characters in the past have affected Stanley's life in the present. These interconnecting stories touch on themes such as racism, homelessness, illiteracy, and arranged marriage. The book was both a

---

Few-shot FT Baseline Response: "Holes by Louis Sachar is a genre of young adult mystery comedy novel [2]. It was first published by Farrar, Straus and Giroux and won the 1998 U.S. National Book Award for Young People's Literature and the 1999 Newbery Medal for the year's m̈ost distinguished contribution to American literature for children[̈1]. The novel centers around an unlucky teenage boy named Stanley Yelnats, who is sent to Camp Green Lake, a juvenile corrections facility in a desert in Texas, after being falsely accused of theft [2]."

CaLF Response: The genre of "Holes" by Louis Sachar is a young adult mystery-comedy [1]. The novel explores themes such as racism, homelessness, illiteracy, and arranged marriage [3].

Figure 5: ASQA Example of generated answers with citations given a question and retrieved passages. We compare the answers produced by the Few-shot FT baseline model with those generated by CaLF.

Question: Why are soap operas instantly identifiable.
Retrieved Passages:
[1] Title: New Soap Chronicles of the Sun To Debut On TV5Monde USA - Soap Opera Spy Text: with Soap Opera Spy for all the latest news, updates and spoilers on all of your favorite daytime and primetime television shows right here! Young and the Restless Sharon Case Talks Mariah-Sharon Drama and Why Nick Is The One Days Of Our Lives Spoilers: JJ Pays It Forward – Talks New Nurse Out Of Suicide, Instantly Smitten With Her

[2] Title: TV Soap Opera - English, Belonging Text: Number 96, which set a new standard and wholly different tone for Australian TV 'soapies'. Other popular and successful local soap operas included The Box (1974), The Young Doctors (1976), The Restless Years (1977), and Prisoner (1979). Teacher to survey the class and lead class discussion: What television soap operas do you watch and why? What soap character do you most closely identify with and why? Viewing questions: Megan Spencer describes the new phase of Neighbours as breaking with the old tradition of "marketing ourselves back to ourselves". How important do you think it is for viewers to see their

[3] Title: Framing The Hobbit – The Golden Lasso Text: per second. What does this mean? Well normal movies and TV shows are shown at 24 frames per second and our minds naturally fill in the missing bits. We don't even realize that it's happening. However, soap operas, home movies, and other videos are shot at 30 frames per second. This is why you can instantly tell a soap opera or reality TV show from other programs just by looking at it for a few seconds. In general, our minds link video and high frame rates to low budget programs. Despite Peter Jackson's assurance that it will revolutionize movies, many

---

Few-shot FT Baseline Response: Soap operas are instantly identifiable because of their distinct tone and style, which is often characterized by a specific storyline and target audience [1]. Additionally, soap operas are easily distinguishable from other programs by their frame rate, which is 30 frames per second, as opposed to the standard 24 frames per second [3].

CaLF Response: Soap operas are instantly identifiable because they are shown at 30 frames per second, which is different from the 24 frames per second used for normal movies and TV shows [3]. Our minds naturally fill in the missing bits, making it easy to tell a soap opera or reality TV show from other programs just by looking at it for a few seconds [3].

Figure 6: ELI5 Example of generated answers with citations given a question and retrieved passages. We compare the answers produced by the Few-shot FT baseline model with those generated by CaLF.

Question: Tell me a bio of Riccardo Tisci.
Retrieved Passages:
[1] Title: Riccardo Tisci Text: Riccardo Tisci Riccardo Tisci (; born 1974) is an Italian fashion designer. He studied in Italy at the Design Istituto d'Arte Applicata in Cantù until the age of 17, and then graduated from London's Central Saint Martins Academy in 1999. In 2005, Tisci was given the title of creative director for Givenchy Women's haute couture and ready-to-wear lines. In May 2008 he was additionally named as menswear and accessories designer of the Givenchy men's division. In March 2018, it was announced he had been appointed chief creative officer of Burberry, succeeding Christopher Bailey. Tisci's apparent fascination with Gothic touches (dark,

[2] Title: Riccardo Tisci Text: rappers Jay-Z and Kanye West to create the cover for "Watch the Throne", as well as the covers for singles H•A•M and Otis. He has now recently collaborated again with Kanye West for his label GOOD Music on their album "Cruel Summer". Riccardo Tisci Riccardo Tisci (; born 1974) is an Italian fashion designer. He studied in Italy at the Design Istituto d'Arte Applicata in Cantù until the age of 17, and then graduated from London's Central Saint Martins Academy in 1999. In 2005, Tisci was given the title of creative director for Givenchy Women's haute couture and ready-to-wear lines.

[3] Title: Riccardo Tisci Text: Tisci worked for companies such as Puma and Coccapani before signing a three-year contract with Ruffo Research, a company that has helped launch the careers of several fashion designers, such as Sophia Kokosalaki. Tisci has mentioned in an interview that he would like to work with an Iranian brand, although it is not known how realistic his wish is. Upon the expiration of his contract in July 2004, Tisci spent time living in India, where he began to work on his own collection. In September 2004, during the Milan Fashion Week, Tisci debuted his first Riccardo Tisci Collection for Fall

---

Few-shot FT Baseline Response: Several years after graduating from London's Central Saint Martins Academy in 1999, Riccardo Tisci was given the title of creative director for Givenchy Women's haute couture and ready-to-wear lines in 2005 [1]. Later, in May 2008, he was additionally named as menswear and accessories designer of the Givenchy men's division [1]. And in March 2018, he was appointed chief creative officer of Burberry, succeeding Christopher Bailey [3]. Tisci's apparent fascination with Gothic touches can be seen in his collaboration with rappers Jay-Z and Kanye West to create the cover for Ẅatch the Throne, as well as the covers for singles H•A•M and Otis [2]. He has also collaborated again with Kanye West for his label GOOD Music on their album C̈ruel Summer[2].

CaLF Response: Riccardo Tisci is an Italian fashion designer, born in 1974 [1]. He studied at the Design Istituto d'Arte Applicata in Cantù in Italy until the age of 17 and later graduated from London's Central Saint Martins Academy in 1999 [1]. Tisci was appointed creative director for Givenchy Women's haute couture and ready-to-wear lines in 2005 [1]. In May 2008, he was also named as menswear and accessories designer for the Givenchy men's division [1]. In March 2018, Tisci was appointed chief creative officer of Burberry, succeeding Christopher Bailey [1]. Tisci has collaborated with rappers Jay-Z and Kanye West to create the cover for "Watch the Throne" and the covers for singles H•A•M and Otis [2]. He has also worked with Kanye West for his label GOOD Music on their album "Cruel Summer" [2].

Figure 7: BIO Example of generated answers with citations given a question and retrieved passages. We compare the answers produced by the Few-shot FT baseline model with those generated by CaLF.