Learning Fine-Grained Grounded Citations for Attributed Large Language Models

Anonymous ACL submission

Abstract

Despite the impressive performance on information-seeking tasks, large language models (LLMs) still struggle with hallucinations. Attributed LLMs, which augment generated text with in-line citations, demonstrate potential in mitigating hallucinations and improving verifiability. Nonetheless, current attributed LLMs suffer from suboptimal citation quality due to their reliance on in-context learning or post-hoc retrieval, lacking a built-in attribution mechanism. Moreover, the practice of merely citing document identifiers falls short in aiding users to pinpoint specific supporting evidence. To bridge these gaps, this work introduces **FRONT**, a training framework that advances the verification process in attributed LLMs through **F**ine-g**R**ained gr**O**u**N**ded ci**T**ations. It equips LLMs with the ability to first anchor in fine-grained supporting quotes, which then guide the generation of attributed answers. Grounded quotes not only elevate LLM attribution quality but also serve as a mechanism for fine-grained verification, significantly enhancing information traceability. Experiments on the ALCE benchmark demonstrate the efficacy of FRONT in generating superior grounded responses and highly supportive citations. With LLaMA-2-7B, the framework significantly outperforms all the baselines, even surpassing ChatGPT, by achieving an average outperformance of 14.21% across all datasets. Notably, FRONT implements an automated procedure and exhibits generalization across models and data scales, enabling continuous performance improvements [1].

1 Introduction

The recent advent of large language models (LLMs) (Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023; OpenAI, 2023, *inter alia*) have taken the world by storm, fueling a paradigm shift in information acquisition

[1]All the data and code will be available soon

Figure 1: Compared with the current attributed systems, the core idea behind FRONT is to first ground the supporting quotes for retrieved documents and then serve as a guide for attributed question answering, ensuring a faithful answer and accurate citation.

(Zhu et al., 2023). Despite their compelling performance, LLMs still struggle with hallucinations (Ji et al., 2023; Zhang et al., 2023; Huang et al., 2023), a tendency to fabricate non-existent facts or generate unfaithful content. This issue further poses a risk of dissemination of misinformation (Chen and Shu, 2023), directly impacting the reliability and trustworthiness of LLMs.

Such prevalence of hallucinations in LLM outputs has motivated the development of attributed systems (Nakano et al., 2021; Thoppilan et al., 2022; Menick et al., 2022), such as New Bing[2] and Perplexity[3], where LLMs are allowed to generate responses with in-line citations. Not only does it

[2]https://www.bing.com/new
[3]https://www.perplexity.ai

improve factuality and alleviate hallucinations, but it also simplifies user verification of model outputs, further enhancing the verifiability of LLMs.

Despite recent advancements, current attributed LLMs still expose significant limitations. **Firstly**, recent efforts in attributed LLMs predominantly rely on either in-context learning (Gao et al., 2023b) or post-hoc retrieval (Gao et al., 2023a), lacking an inherent capability for attributable generation, thereby resulting in compromised citation quality (Liu et al., 2023b). **Secondly**, current attributed systems typically cite either document identifiers (Nakano et al., 2021) or URLs (Thoppilan et al., 2022), which complicates the process for users to pinpoint the exact supporting quotes, particularly in lengthy documents.

To this end, we explore *how to empower LLMs to learn a built-in attribution mechanism while providing fine-grained verification*. Recognizing the inherent of verification lies in grounding, we utilize it as the bridge between verification and attribution. By anchoring generated content to fine-grained grounded quotes, attribution is seamlessly integrated. Consequently, we propose a unified framework FRONT, designed to advance coarse verification via **F**ine-g**R**ained gr**O**u**N**ded ci**T**ations, concurrently enhancing attributability. Specifically, FRONT starts with a pipeline tailored for the automated generation of high-quality, attributed data, serving as the supervised signals for effectively injecting attributability. Given a user query, the pipeline automates data construction through document retrieval, reranking, attributed answer generation, and data filtering to ensure the informativeness and attributability of the answers. Furthermore, to unlock LLMs' ability for attributable generation while providing fine-grained verification, we propose a two-stage training recipe, Grounding Guided Generation ($\boldsymbol{G^3}$) and Weak-to-Strong Contrastive Alignment (**CA**) (§4.2). $\boldsymbol{G^3}$ equips the model with the ability to first anchor in fine-grained supporting quotes, which then guide the generation of attributed answers. While **CA** further improves the consistency of grounded quotes and attributed answers by automatically constructing contrastive signals from weak and strong LLMs.

We conduct extensive experiments to evaluate our framework on the ALCE Benchmark (Gao et al., 2023b). Findings are: (1) Training on high-quality synthetic data markedly boosts citation quality. With LLaMA-2-7B, FRONT significantly surpasses ChatGPT, achieving an average outperformance of 14.21%. (2) FRONT demonstrates remarkable generalization across models and data scales. (3) Abalation studies confirm the significance of each component and underscore the potential of FRONT for continuous performance improvements.

2 Related Work

Retrieval Augmented Generation. Recently, Retrieval Augmented Generation (RAG) has shown promise in knowledge-intensive tasks by incorporating retrieved documents into LLM input, equipping models with pertinent knowledge to reduce hallucinations (Karpukhin et al., 2020; Lewis et al., 2020; Feng et al., 2023; Gao et al., 2023c). Despite these advancements, recent studies have identified challenges in handling irrelevant or contradictory documents (Shi et al., 2023; Yoran et al., 2023; Xu et al., 2023) and effectively utilizing input context (Liu et al., 2023a), underscoring the necessity for more factual and verifiable systems.

Attributed Large Language Models. The persistent challenge of hallucinations within LLMs has spurred the development of attributed LLMs (Bohnet et al., 2022; Li et al., 2023a; Worledge et al., 2023), which seek to enhance information verifiability by generating responses with in-line citations. The way of providing attributions varies across studies. For example, Gao et al. (2023b) enables LLMs to generate text with in-line citations via in-context learning. Gao et al. (2023a) explores post-hoc attribution, where LLMs first generate an initial response and then retrieve relevant evidence to achieve attribution. Furthermore, Li et al. (2023b); Asai et al. (2023); Sun et al. (2023) explores adaptive retrieval for attribution, where a verifier provides feedback for flexible retrievals.

3 Preliminaries: Task Formulation

Following (Liu et al., 2023b; Gao et al., 2023b), the task is formalized as follows: Given a user query q and a corpus of retrieved documents $\mathcal{D}$ as input, the LLM is required to produce a response $\mathcal{S}$, which consists of statements with embedded in-line citations. We assume the response $\mathcal{S}$ comprising with n statements $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and each statement $s_i \in \mathcal{S}$, cites a list of passage $\mathcal{C}_i = \{c_{i1}, c_{i2}, \ldots\}$, where $c_{ij} \in \mathcal{D}$. Specifically, citations are presented in the form of `[1][2]`.

4 Methodology

This section outlines FRONT, which comprises two components, as illustrated in Figure 2: an automatic data generation pipeline(§4.1) and a two-stage training recipe (§4.2).

4.1 Automatic Data Generation Pipeline

Equipping LLMs with built-in attribution capabilities requires training data consisting of elaborate responses paired with accurate citations, which typically requires a labor-intensive and costly manual process. To address this, we propose a pipeline designed for the automatic generation of high-quality attributed data, encompassing data collection, attributed answer generation, and data filtering.

Data Collection. To simulate the real-world environment for information-seeking, we source questions from the Natural Question (NQ) (Kwiatkowski et al., 2019) dataset, which consists of real user queries from the Google search engine. The dataset spans a range of diverse question types, demanding answers of varying lengths, from concise to detailed. To mimic the way a search engine might synthesize documents of high relevance in response to a user query, we employ Sphere (Piktus et al., 2021), a pre-processed and cleaned version of the Common Crawl corpus, serving as a proxy web search index. In particular, for a given user query sampled from the NQ dataset, we initially retrieve the top-100 relevant documents from the Sphere corpus using sparse retrieval. These documents are subsequently re-ranked by RankVicuna (Pradeep et al., 2023) considering the superior performance in listwise re-ranking, resulting in the top-5 most relevant documents for each query.

Attributed Answer Generation. Given the remarkable performance of ChatGPT in attributed question answering, we employ ChatGPT to generate answers with corresponding citations for given queries and the top-5 retrieved documents. We provide precise instructions and in-context demonstrations to ensure that ChatGPT produces informative responses and cites the sources accordingly.

Data Filtering. To guarantee the high quality of our synthetic training data, we employ a data filtering process guided by two key criteria derived from Kamalloo et al. (2023): (1) *informativeness*: assessing if the answer provides sufficient information to the question, and (2) *attributability*: determining if the answer is attributed to the cited documents. To mitigate the impact of nonsensical queries and irrelevant document retrieval that may lead to non-informative answers, we utilize ChatGPT for preliminary informativeness annotations. Responses categorized as non-informative are directly excluded. Furthermore, to ensure that answers are accompanied by highly supportive citations, we train a discriminator on human-labeled data from the comprehensive evaluation by Liu et al. (2023b), where attributability is categorized into three levels: full support, partial support, or no support. We quantitatively map the discriminator's outputs to an attributability score and ultimately derive an average score for each attributed answer. Answers falling below a defined threshold are systematically excluded to ensure the synthetic data's reliability, which results in nearly 8,000 entries. For more details, please refer to Appendix A.1.

4.2 Two-Stage Training Recipe

To equip LLMs with built-in attribution capabilities while ensuring fine-grained verification, we introduce a two-stage training recipe, which consists of two steps: 1) **G**rounding **G**uided **G**eneration (**G^3**), designed for unlocking the ability to generate grounded quotes then guide attributed answer generation; 2) Weak-to-Strong **C**ontrastive **A**lignment (**CA**), aimed at enhancing the consistency between grounding and attributed answers while facilitating precise and supportive citations by contrastive disparities from weak and strong LLMs.

4.2.1 Grounding Guided Generation

To empower LLMs with attribution capabilities while ensuring fine-grained verification, we propose **G**rounding **G**uided **G**eneration (**G^3**), which employs grounding as a crucial bridge between verification and attribution. The cornerstone of **G^3** lies in enabling LLMs to extract supporting quotes from the source documents, each associated with its document identifier, which in turn guides the generation of attributed answers. Such a grounding format offers two primary benefits: Firstly, the direct extraction of quotes from sources significantly reduces the impact of the incorporation of irrelevant information and the risk of hallucinations in subsequent attributed answers. Secondly, the process naturally facilitates accurate attribution, with each document identifier serving as a clear supervised signal that delineates the origin of the extractive quotes, thus improving the citation quality.

However, the absence of specific grounding con-

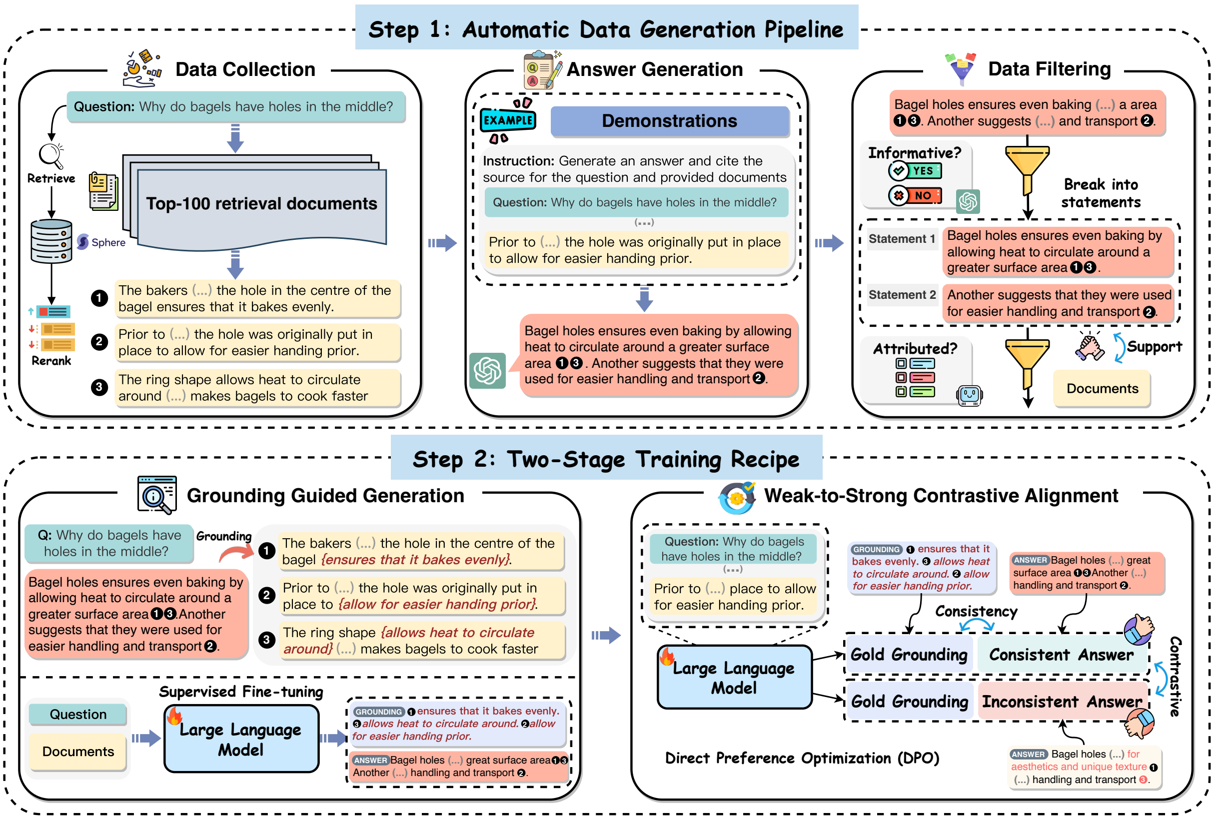

Figure 2: Overview of FRONT: The Data Generation module facilitates the automatic generation of diverse and high-quality attributed answers. The two-stage training recipe then enables LLMs to first generate precise grounding and subsequently guide the generation of attributed answers, thereby enhancing fine-grained verification capabilities.

tent for statements within our synthetic attributed answers poses additional challenges. To tackle this, we employ ChatGPT to meticulously extract segments from cited documents that support the corresponding statement. Hence, when given a query q and top-5 retrieved documents $\mathcal{D}$ as input, the LLM is fine-tuned to generate a response $\mathcal{S}$ which consists of two components: the extractive grounding $\mathcal{G}$ and the attributed answer $\mathcal{A}$. Specifically, the extractive grounding $\mathcal{G}$ is delineated as follows:

$$\mathcal{G} = \{\texttt{[GROUNDING]}, (i_1, e_1), \ldots, (i_n, e_n)\}, \quad (1)$$

where [GROUNDING] denotes a special token indicating the start of grounding content. Each tuple within $\mathcal{G}$, comprising a document identifier i and the corresponding extractive segment e, collectively forming an extractive grounding quote.

Similarly, the formulation of the attributed answer $\mathcal{A}$ is concisely presented as:

$$\mathcal{A} = \{\texttt{[ANSWER]}, s_1, s_2 \ldots, s_m\}, \quad (2)$$

where [ANSWER] is a special token that signals the beginning of the attributed answer. Each statement s_i cites a list of passages $\mathcal{C}_i = \{c_{i1}, c_{i2}, \ldots\}$, where $c_{ij} \subseteq \{i_1, i_2, \ldots, i_n\}$, as defined in Equation 2.

Thus, the training loss is formulated as:

$$\mathcal{L} = -\sum_{i=1}^{N} \log P(y_i | q_i, \mathcal{D}_i; \theta) \quad (3)$$

where y_i represents the combined output of grounding $\mathcal{G}$ and answer $\mathcal{A}$ for each given query q_i and set of retrieved documents $\mathcal{D}_i$.

4.2.2 Weak-to-Strong Contrastive Alignment

While $\boldsymbol{G^3}$ unlocks the ability to extract supporting quotes followed by generating attributed answers, it occasionally leads to inconsistencies between grounding quotes and attributed answers. Such discrepancies challenge the attempt to employ these grounding quotes as fine-grained verification. In response, we propose a contrastive-based alignment stage specifically aimed at enhancing the consistency between grounding and answer generation.

The cornerstone of our approach involves contrasting a consistent answer with an inconsistent one under the guidance of the same oracle grounding quotes, which aligns with the concept of

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022), where LLMs are further fine-tuned to distinguish between desirable and undesirable responses under preference feedback. However, such contrastive preference feedback typically comes from human annotation. Inspired by the *weak-to-strong generalization* (Burns et al., 2023; Zhao et al., 2024) where a weaker LLM is utilized to guide the training of more powerful LLMs, we introduce weak-to-strong contrastive alignment (**CA**) that employs smaller LLMs (*e.g.*, 7B) to provide contrastive supervision signals. In this setting, the process not only encourages the LLM to generate attributed answers more consistent with the grounding quotes but also facilitates the identification and correction of nuanced errors present in smaller models.

Specifically, we adopt Direct Preference Optimization (Rafailov et al., 2023), a variant of RLHF known for its stability, for our contrastive alignment. Formally, for each instance, given the oracle grounding $g^{(i)}$ along with a consistent oracle answer $y_w^{(i)}$ as well as an attributed answer $y_l^{(i)}$ generated by a weaker LLM via in-context learning, we can simply construct a preference dataset:

$$\mathcal{D} = \{x^{(i)}, \tau_w^{(i)}, \tau_l^{(i)}\}_{i=1}^{N}, \tag{4}$$

where $\tau_w^{(i)} = g^{(i)} \circ y_w^{(i)}$ denotes the concatenation of the oracle grounding with the consistent, attributed answer, $\tau_l^{(i)} = g^{(i)} \circ y_l^{(i)}$ denotes the concatenation with the inconsistent attributed answer. Here, $\circ$ signifies the operation of string concatenation.

Finally, we can optimize the policy model π_θ on the dataset $\mathcal{D}$ by minimizing the following loss:

$$\begin{aligned} &\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{ref}; \mathcal{D}) \\ &= -\mathbb{E}_{(x,\tau_w,\tau_l)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta(\tau_w|x)}{\pi_{\text{ref}}(\tau_w|x)} \\ &\quad -\beta\log\frac{\pi_\theta(\tau_l|x)}{\pi_{\text{ref}}(\tau_l|x)}\Bigg)\Bigg], \end{aligned} \tag{5}$$

where π_{ref} represents the reference model, initialized from $\boldsymbol{G^3}$. The hyper-parameter β modulates the divergence between the distribution from the policy model and the reference model. τ_w is the consistent answer, while τ_l is the inconsistent one.

5 Experimental Settings

5.1 Datasets

We conduct experiments on the ALCE benchmark (Gao et al., 2023b), offering a collection of diverse datasets spanning various question types and a comprehensive suite for automatic evaluation of LLM attribution which exhibits a strong correlation with human judgments. The benchmark includes:

ASQA (Stelmakh et al., 2022) is a long-form factoid QA dataset characterized by questions that are inherently ambiguous, necessitating multiple short answers to encapsulate different viewpoints.

ELI5 (Fan et al., 2019) is a long-form QA dataset that features open-ended questions requiring explanatory multi-sentence answers.

QAMPARI (Amouyal et al., 2022) is a factoid QA dataset derived from Wikipedia, where answers are structured as a compilation of entities.

5.2 Evaluation Metrics

Following the ALCE benchmark (Gao et al., 2023b), our evaluation primarily focuses on two key dimensions. A more comprehensive evaluation is presented in the Appendix E.

Citation Quality. Citation quality is critical for evaluating LLM attribution, assessed along two dimensions: (1) *Citation Recall*, determining if the output is entirely supported by the cited documents, and (2) *Citation Precision*, assessing if each citation supports its corresponding statement. Evaluation is conducted by TRUE (Honovich et al., 2022), a T5-11B model fine-tuned on a collection of NLI datasets to automatically examine the entailment of cited documents and the model generation. Additionally, to capture a holistic measure of citation quality, we also report the *Citation F1*, the harmonic mean of citation precision and recall:

$$F_1 = 2 \cdot \frac{\text{citation precision} \cdot \text{citation recall}}{\text{citation precision} + \text{citation recall}}, \tag{6}$$

Correctness. Correctness is determined by comparing the accuracy of responses to ground truth answers. For ASQA, correctness is quantified using EM recall to capture the recall of correct short answers. Regarding ELI5, correctness is measured by the entailment between ground truth sub-claims and the model's response. For QAMPARI, the model's generation correctness is assessed through both top-5 EM recall and EM precision.

5.3 Baselines

We compare our method with three types of baselines: prompting-based, post-hoc retrieval, and training-based approach.

Model Type	Model Size	ASQA				ELI5				QAMPARI				
		Correctness	Citation			Correctness	Citation			Correctness		Citation		
		EM Rec.	Rec.	Prec.	F1.	Claim	Rec.	Prec.	F1	Rec.-5	Prec.	Rec.	Prec.	F1
Prompting-based														
ChatGPT	-	40.37	72.81	69.69	71.22	12.47	49.44	47.05	48.22	20.28	19.84	19.06	22.03	20.44
LLaMA-2	7B	24.32	17.24	17.87	17.55	4.53	3.92	5.38	4.54	12.56	11.32	6.03	6.35	6.19
	13B	27.99	16.45	19.04	17.65	7.77	8.49	8.43	8.46	18.00	12.39	5.45	5.74	5.59
	70B	31.53	44.18	44.79	44.48	10.43	23.75	22.43	23.07	18.50	14.79	10.10	10.50	10.30
LLaMA-2-Chat	7B	29.93	55.99	51.66	53.74	12.47	19.90	15.48	17.41	17.96	19.74	9.58	9.68	9.63
	13B	34.39	37.15	38.17	37.65	13.83	16.50	16.09	16.29	21.34	18.86	8.94	9.06	9.00
	70B	41.24	60.19	61.16	60.67	13.30	36.63	36.63	36.63	22.62	18.04	13.49	13.98	13.73
Vicuna-v1.5	7B	38.34	48.37	44.63	46.42	12.30	29.81	22.45	25.61	14.22	14.74	11.26	11.64	11.45
	13B	35.20	51.92	53.40	52.65	14.33	31.15	28.99	30.03	22.06	19.60	13.04	13.74	13.38
Mistral	7B	29.46	23.12	25.45	24.23	8.47	16.04	16.32	16.18	16.96	15.98	7.50	7.76	7.63
	8 × 7B	36.30	32.72	34.49	33.58	10.43	26.11	25.09	25.59	18.18	15.63	9.72	10.20	9.95
Mistral-Instruct	7B	38.57	64.90	59.67	62.18	11.07	49.25	42.69	45.74	17.52	21.29	17.56	18.53	18.03
	8 × 7B	**44.11**	61.80	63.27	62.53	13.93	49.28	48.34	48.81	20.12	19.64	19.27	20.38	19.81
Post-hoc Retrieval-based														
ChatGPT	-	37.68	27.11	27.05	27.08	**18.77**	14.55	14.55	14.55	**25.14**	**22.85**	12.29	12.29	12.29
LLaMA-2-Chat	70B	29.68	24.51	24.51	24.51	16.03	12.93	12.93	12.93	17.90	14.45	9.05	9.05	9.05
Mistral-Instruct	8 × 7B	33.90	24.57	24.48	24.52	17.37	15.68	15.68	15.68	24.16	18.28	9.78	9.78	9.78
Training-based														
Self-RAG (LLaMA-2)	7B	29.96	66.97	67.82	67.39	-	-	-	-	-	-	-	-	-
	13B	31.53	58.32	68.35	62.94	-	-	-	-	-	-	-	-	-
VANILLA-SFT (LLaMA-2)	7B	40.32	67.67	63.67	65.61	9.63	42.30	40.06	41.15	12.86	21.09	21.35	21.36	21.35
	13B	40.85	71.49	66.21	68.75	10.27	46.75	44.47	45.58	12.68	22.80	23.64	23.71	23.67
FRONT (LLaMA-2)	7B	40.84	77.70	69.89	73.59	10.18	58.60	55.33	56.92	11.50	21.38	24.74	24.84	24.79
	13B	41.51	**78.44**	**73.66**	**75.97**	10.32	**60.31**	**59.21**	**59.75**	11.94	22.61	**24.86**	**25.39**	**25.12**

Table 1: Main results on the ALCE benchmark. **Bold** numbers indicate the best performance, while _ indicates the second-best performance. - indicates numbers that are not applicable.

5.3.1 Prompting-based Methods.

Our experiments span a spectrum of LLMs, ranging from foundational models to supervised fine-tuning (SFT) LLMs. For foundational LLMs, we select **GPT-3.5-Turbo**[4] as the representative closed-source model for its notable performance. Among the open-source foundational LLMs, we focus on the LLaMA-2 series including **LLaMA2-7B**, **LLaMA2-13B**, and **LLaMA2-70B**, as well as the Mistral series, which spans from **Mistral-7B** to **Mistral-8x7B-MoE**. Regarding SFT LLMs, we select the SFT counterparts of the open-source foundational LLMs we used. Detailed prompting settings can be found in Appendix B.

5.3.2 Post-hoc Retrieval-based Methods.

Following Gao et al. (2023b), we apply the previously mentioned models to perform post-hoc retrieval. Initially, LLMs are prompted to generate answers in a closed-book setting. Subsequently, for each generated statement, we utilize GTR to identify and cite the most relevant document from the top-100 retrieved documents.

5.3.3 Training-based Methods.

Self-RAG (Asai et al., 2023) trains the LLM to learn to adaptively retrieve passages on-demand and enable it to reflect on its generation to further improve generation quality and attributions.

VANILLA-SFT trains the LLM directly on synthetic training data, where, given a query and retrieved documents, the LLM are learnt to directly generates attributed answers.

[4]Specifically, we utilize `gpt-3.5-turbo-1106` version

5.4 Implement Details

We conduct FRONT with different foundational models to evaluate its effectiveness: LLaMA-2-7B and LLaMA-2-13B. The comprehensive training details are presented in Appendix C.2

6 Results and Analysis

6.1 Overall Results

Training LLMs to equip built-in attribution ability boosts citation quality. As shown in Table 1, equipping LLMs with the capability for attribution through training markedly boosts citation quality, showing significant advancements over both prompt-based and post-hoc baselines across all datasets. Specifically, simply supervised fine-tuning (VANILLA-SFT) on our synthetic data with the LLaMA-2-7B model led to substantial gains in citation F1 scores over prompting: ASQA (17.55 → 65.61), ELI5 (4.54 → 41.15), and QAMPARI

Model	ASQA				ELI5				QAMPARI				
	Correctness	Citation			Correctness	Citation			Correctness		Citation		
	EM Rec.	Rec.	Prec.	F1.	Claim	Rec.	Prec.	F1	Rec.-5	Prec.	Rec.	Prec.	F1
FRONT-7B	40.84	77.70	69.89	73.59	10.18	58.60	55.33	56.92	11.50	21.38	24.74	24.84	24.79
SELF-GUIDE (w/o Contrastive)	38.99	70.69	64.48	67.44	10.04	47.63	44.80	46.17	12.18	20.03	22.50	22.58	22.54
VANILLA-SFT (w/o Ground)	40.32	67.67	63.67	65.61	9.63	42.30	40.06	41.15	12.86	21.09	21.35	21.36	21.35
FRONT-13B	41.51	78.44	73.66	75.97	10.32	60.31	59.21	59.75	11.94	22.61	24.86	25.39	25.12
SELF-GUIDE (w/o Contrastive)	40.99	73.08	68.13	70.52	10.06	50.68	49.78	50.23	13.94	22.38	23.73	23.99	23.85
VANILLA-SFT (w/o Ground)	40.85	71.49	66.21	68.75	10.27	46.75	44.47	45.58	12.68	22.80	23.64	23.71	23.67

Table 2: Ablation study on the impact of different training stages within the ALCE benchmark.

(6.19 → 21.35), which also highlight the efficacy of the synthetic data generation procedure in FRONT.

FRONT achieves significant performance gains and surpasses ChatGPT. While VANILLA-SFT achieves strong performance, notable disparities still exist compared to leading open-source LLMs, such as Mixtral-8×7B-Instruct (*e.g.*, 41.15 vs. 45.74) and ChatGPT (*e.g.*, 41.15 vs. 48.22) on the ELI5 dataset. FRONT not only narrows these gaps but also establishes significant leads across all datasets. Specifically, with LLaMA-2-7B, FRONT achieves comprehensive outperformance over ChatGPT, registering outperformance of 3.32%, 18.04%, and 21.28% on the ASQA, ELI5, and QAMPARI datasets respectively, which underscores the effectiveness of FRONT in enhancing attribution capabilities.

FRONT exhibits scalability with model size. As illustrated at the bottom of Table 1, the performance of FRONT shows significant enhancements when transitioning from 7B to 13B, with improvements of 3.23%, 4.97%, and 1.33% on the ASQA, ELI5, and QAMPARI, respectively. This upward trend underscores the scalability of FRONT with model size, demonstrating the potential of FRONT in leveraging the increased capabilities of larger LLMs for further performance gains.

FRONT demonstrates remarkable generalization and improves correctness. Compared to the varied queries and answer types present in the ALCE, our synthetic training data, derived exclusively from the NQ dataset, exhibits out-of-domain characteristics. Nonetheless, FRONT demonstrates superior citation quality, affirming its exceptional ability to generalize across queries. Furthermore, although FRONT was not explicitly designed to optimize for correctness, it showcases notable improvements over VANILLA-SFT, particularly achieving significant performance advancements on the ASQA and QAMPARI datasets, and even outperforming ChatGPT. However, FRONT encounters lower citation recall on the QAMPARI dataset, likely due to its answers, which are composed of concatenated entities, significantly diverging from our training data's distribution.

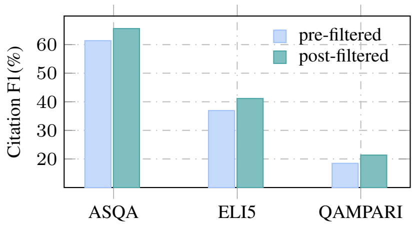

Figure 3: Ablation Study on Data Filtering.

6.2 Ablation Study

We conduct ablation studies to verify the effectiveness of different components proposed in FRONT.

Effects of Data Generation Pipeline. As illustrated in §6.1, simply SFT achieves strong performance, underscoring the high quality of our synthetic data. Furthermore, data filtering, a crucial component of our data generation pipeline, plays a pivotal role in ensuring the quality of the generated data by filtering out queries that yield non-informative answers or fail to meet attribution criteria. To validate the effectiveness of our data filtering strategies, we conducted experiments comparing models fine-tuned on both pre-filtered and post-filtered data. The results, depicted in Figure 3, confirm that models trained on filtered data exhibit a notable improvement in citation quality over those trained on unfiltered data, achieving superior attribution performance with reduced data volume.

Effects of Grounding Guided Generation. To validate the effectiveness of Grounding Guided Generation (G^3) in improving attribution, we compared the model, SELF-GUIDE, trained solely through the G^3 stage (w/o Contrastive) against VANILLA-SFT (w/o Ground), which is trained with the same synthetic data but allowed to generate

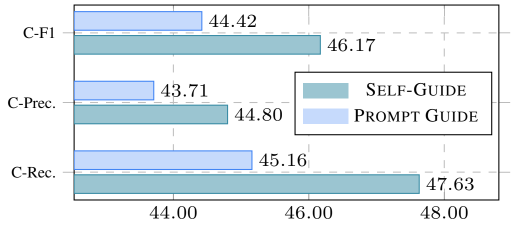

Figure 4: Ablation study of different grounding guidance forms on the ELI5 dataset.

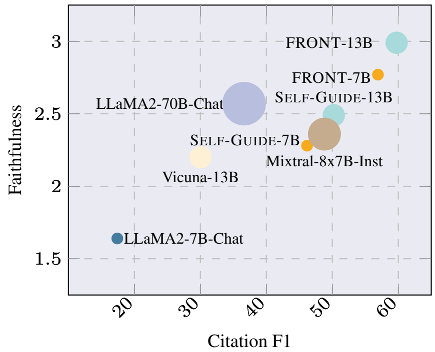

Figure 5: The relationship between citation F1 and hallucination: Models positioned closer to the top-right corner exhibit higher citation quality and a lower degree of hallucination.

attributed answers directly, bypassing the grounding step. The ablation study, detailed in Table 2, reveals that models incorporating grounding guidance markedly surpass their VANILLA-SFT counterparts, which lack such grounding mechanisms. This underscores the pivotal role of grounding in bolstering attribution.

Moreover, we explore an alternative variant of grounding guidance, PROMPT-GUIDE, trained to leverage grounding guidance within the prompt along with the query and retrieved documents for generating attributed answers. During inference, PROMPT-GUIDE employs oracle grounding content, extracted by ChatGPT, to incorporate grounding guidance. Conducting experiments on the ELI5 dataset using the LLaMA-2-7B model, the results depicted in Figure 3 reveal that SELF-GUIDE outperforms PROMPT-GUIDE. This finding underscores that training models to generate grounding before attributed answers yields more effective results than merely using grounding as prompt guidance, highlighting the superiority of FRONT.

Effects of Weak-to-Strong Alignment (CA). The goal of **CA** is to enhance the consistency between grounded quotes and attributed answers, thereby alleviating hallucinations and achieving more precise attribution. To this end, we compared models that underwent only the G^3 stage (SELF-GUIDE) with the one further enhanced through the CA stage (FRONT). As illustrated in Table2, FRONT significantly improves citation quality over SELF-GUIDE, demonstrating the effectiveness of the CA stage in enhancing attribution.

Moreover, to assess the impact of **CA** on reducing hallucinations, we employed QAFactEval, a QA-based factual consistency metric measuring the consistency between model responses and given documents. Specifically, we analyzed the performance of leading open-source models and two variants of FRONT and SELF-GUIDE on the ELI5 dataset. As shown in Figure5, FRONT produces more faithful outputs than SELF-GUIDE, significantly reducing hallucinations.

Effects of Training Data Scale. We analyze the impact of the data scale on model performance across two training stages. In particular, we randomly sampled 2k, 4k, 6k, and 8k instances from our full training data across two distinct training stages. These subsets were then utilized to fine-tune various 7B model variants, enabling a comparative analysis of performance based on data scale. Results are shown in Figure 6, which indicates that increasing data size shows significant enhancements in citation quality, indicating a positive correlation between data size and model performance. As FRONT implements an automated procedure capable of generating high-quality attributed data and constructing contrastive supervision from weak and strong LLMs, it holds the potential for continuous performance improvements.

7 Conclusion

In this work, we introduce FRONT, a training framework designed to enpower LLMs with the capability for built-in attribution while facilitating fine-grained verification. FRONT encompasses an automated data generation pipeline, crafting high-quality synthetic data that trains an LLM to first generate grounded quotes and then subsequently guide the generation of attributed answers. Notably, by enhancing the consistency between the grounding and attributed answers, FRONT takes a significant leap forward, harnessing grounding as a mechanism for fine-grained verification. Through comprehensive experiments, FRONT has been shown to produce superior grounded responses and highly supportive citations, significantly outperforming existing methods, even surpassing ChatGPT.

8 Limitation

Our study presents several limitations worth noting. Firstly, the validation of our framework is predominantly conducted on models of sizes 7B and 13B, leaving the exploration of larger models, such as LLaMA2-70B due to computational constraints. Secondly, our framework relies on a prior retrieval process, wherein relevant documents are retrieved at one time. The incorporation of adaptive retrieval, enabling more dynamic interactions with LLMs, could potentially enhance performance. We leave it for future research. Lastly, evaluating the correctness of long-form question answering presents inherent challenges, leading our framework to primarily enhance citation quality, with modest advancements in correctness. Therefore, we advocate for the development of more robust metrics capable of accurately assessing the correctness of long-form QA responses, paving the way for future work.

Acknowledgements

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Samuel Joseph Amouyal, Ohad Rubin, Ori Yoran, Tomer Wolfson, Jonathan Herzig, and Jonathan Berant. 2022. QAMPARI: : An open-domain question answering benchmark for questions with many answers from multiple paragraphs. *CoRR*, abs/2205.12665.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *CoRR*, abs/2212.08037.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *CoRR*, abs/2312.09390.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *CoRR*, abs/2311.05656.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek

Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics.

Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications. *CoRR*, abs/2311.05876.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16477–16508. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *CoRR*, abs/2305.14627.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023c. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3905–3920. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *CoRR*, abs/2010.12694.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023a. A survey of large language models attribution. *CoRR*, abs/2311.03731.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023b. Llatrieval: Llm-verified retrieval for verifiable generation. *CoRR*, abs/2311.07838.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023b. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7001–7025. Association for Computational Linguistics.

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

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick S. H. Lewis, Barlas Oguz, Edouard Grave, Wen-tau Yih, and Sebastian Riedel. 2021. The web is your oyster - knowledge-intensive NLP against a very large web corpus. *CoRR*, abs/2112.09924.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *CoRR*, abs/2309.15088.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8273–8288. Association for Computational Linguistics.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2023. Towards verifiable text generation with evolving memory and self-reflection. *CoRR*, abs/2312.09075.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. 2023. Unifying corroborative and contributive attributions in large language models. *CoRR*, abs/2311.12233.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. RECOMP: improving retrieval-augmented lms with compression and selective augmentation. *CoRR*, abs/2310.04408.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *CoRR*, abs/2310.01558.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. 2024. Weak-to-strong jailbreaking on large language models.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.

A Details of Data Generation

A.1 Data Statistic

# Questions	8,098
➥ # Long Answer	5667
➥ # Short Answer	2431
Avg. Words per Answer	50.48
➥ Avg. Words per Long Answer	69.15
➥ Avg. Words per Short Answer	6.94
Avg. Citation per Answer	4.40
➥ Avg. Citation per Long Answer	4.68
➥ Avg. Citation per Short Answer	3.77

Table 3: The statistics of the data generated by our automatic data generation pipeline.

Table 3 presents the statistics of the data automatically generated by our data generation pipeline. In total, we collected 8,098 questions from the Natural Questions (NQ) dataset, of which 5,667 questions were gathered from those with long-form answers, and 2,431 questions were collected from those with short-form factoid answers.

For questions requiring long-form answers, we initialized our query source with the AQUAMUSE dataset (Kulkarni et al., 2020), which consists of high-quality queries specifically designed for long-form responses within the NQ dataset, recognized as "good" by the majority of NQ evaluators. In this way, utilizing a refined and superior quality query set laid a robust groundwork for our training data generation, streamlining the data filtering process. For factoid queries that necessitate short-form answers, we directly sampled from the original NQ dataset, leveraging its abundance and inherently high quality.

During the data generation process, our initial query set comprised 7,725 queries requiring long-form answers and 4,000 queries necessitating short-form answers. After a two-stage data filtering process, we retained 5,667 and 2,431 queries, respectively. Additionally, we calculated the average length of answers and the average number of citations generated for various types of queries within our dataset, as shown in Table 3.

B Prompts

B.1 Prompts for Prompting-Based Methods

Following Gao et al. (2023b), we adopt the vanilla prompting strategy for its simplicity and effectiveness. Specifically, the prompts vary according to the type of data within the ALCE benchmark. For long-form QA datasets such as ASQA and ELI5, the prompt format is detailed in Table 4. For the short-form QA dataset QAMPARI, the format is outlined in Table 5.

B.2 Instructions for FRONT

During the training process, we follow the instruction format of Alpaca[5]. Specifically, we employ varied instructions for different question types, as delineated in Table 6 for long-form questions and Table 7 for short-form questions.

C Experimental Details

C.1 More Details of Attributed Discriminator

We trained our Attributed Discriminator using the manually annotated data provided by Liu et al. (2023b), which is sampled from real generative search engines. Each statement and its cited document have been meticulously annotated for attribution, categorized into three types: complete support, partial support, and no support. For training, we utilized a dataset of 8,834 instances, comprising 6,415 instances of complete support, 1,552 of partial support, and 867 of no support. The discriminator initialized with LLaMA-2-7B, was trained with a maximum sequence length of 512. We trained it for 3 epochs, with a total batch size of 128, and a peak learning rate of 2e-5, incorporating 3% warmup steps, followed by a linear decay.

During the data filtering stage, we first break down the automatically generated attributed answers into statement form and use the trained discriminator to annotate the attribution between each statement and its cited documents. Specifically, we assign different attribution scores to each statement s based on its attribution relationship with cited documents d, as shown in Equation 7. Consequently, for each attributed answer, we can calculate its average attribution score. Attributed answers with an average attribution score below 0.8 are filtered out. The threshold of 0.8 was determined through preliminary testing on the development set, for which we manually annotated 100 samples to ensure the effectiveness of our filtering criteria.

$$r(s) = \begin{cases} 1, & \texttt{Dis}(s,d) = \texttt{complete support} \\ 0.5, & \texttt{Dis}(s,d) = \texttt{partial support} \\ 0, & \texttt{Dis}(s,d) = \texttt{no support} \end{cases} \tag{7}$$

C.2 More Details of Training in FRONT

The training of all models is executed on 4 Nvidia A100 GPUs, each with 80GB of memory, leveraging the Deepspeed (Rasley et al., 2020) and HuggingFace Accelerate libraries (Gugger et al., 2022) to conduct multi-GPU distributed training. Given the long nature of the inputs, the maximum token length is set to 2,048 tokens.

During the grounding guide generation stage, models are trained for 5 epochs with a total batch size of 128, a peak learning rate of 2e-5 with 3% warmup steps followed by a linear decay. During the contrastive alignment stage, we set the β to 0.1 and continued training for two additional epochs. Specifically, During inference, we use the vllm framework (Kwon et al., 2023) for efficient inference. The hyperparameters are set as illustrated in Table 8.

D More detail about Ablation Study

D.1 The Effect of Training Data Scale.

We examine how model performance varies with changes in data scale, as depicted in Figure6. The upper part of the figure illustrates the impact of the training data scale on citation quality during the Grounding Guided Generation training stage, with datasets ASQA, ELI5, and QAMPARI represented from left to right. Similarly, the lower part of the figure describes the influence during the Weak-to-Strong Alignment training stage.

D.2 The Generalization Across Model Architectures.

FRONT demonstrates exceptional generalization capabilities across various foundational model architectures. Specifically, transitioning the foundational model from LLaMA-2-7B to the stronger foundational model, Mistral-7B, results in even greater performance enhancements as shown in Figure 7. This further underscores the broad applicability and generalizability of FRONT.

[5] https://github.com/tatsu-lab/stanford_alpaca/tree/main

```
Instruction: Write an accurate, engaging, and concise answer for the given question using only
the provided search results (some of which might be irrelevant) and cite them properly. Use an
unbiased and journalistic tone. Always cite for any factual claim. When citing several search
results, use [1][2][3]. Cite at least one document and at most three documents in each sentence.
If multiple documents support the sentence, only cite a minimum sufficient subset of the
documents.
```

Table 4: Prompt for Long-form QA.

```
Instruction: Provide a list of accurate answers for the given question using only the provided
search results (some of which might be irrelevant) and cite them properly. Always cite one and
only one document for each answer. Separate answers by commas. For questions that have more than
5 answers, write at least 5 answers.
```

Table 5: Prompt for Short-form QA.

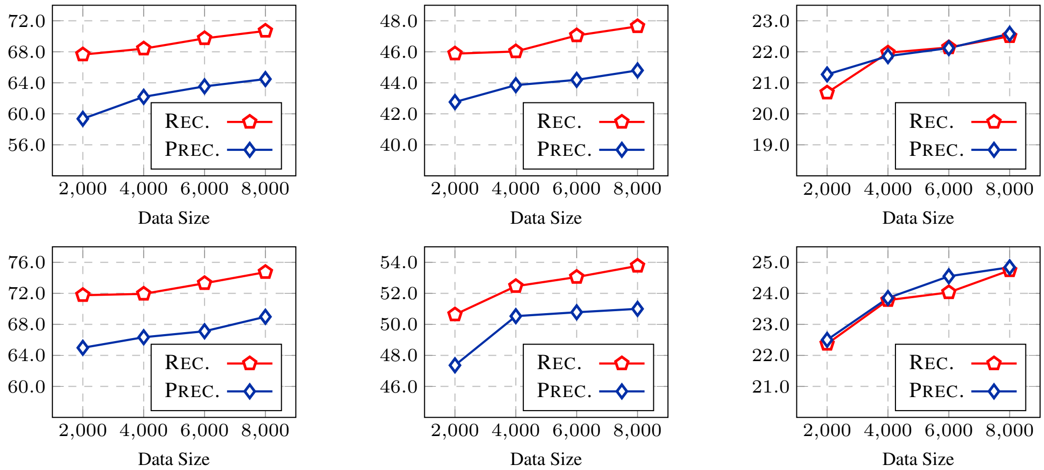

Figure 6: Ablation study on synthetic training data size: The upper part of the figure corresponds to the Grounding Guided Generation training stage, while the bottom part represents the Weak-to-Strong Contrastive Alignment training stage. From left to right, the results are presented for ASQA, ELI5, and QAMPARI, respectively. REC. indicates Citation Recall and PREC. denotes Citation Precision. The x-axis represents the quantity of automatically generated data. It is observed that as the volume of automatically generated data increases, there is a consistent improvement in both citation recall and precision across the two training stages.

D.3 The effect of β in Weak-to-Strong Contrastive Alignment Training Stage

In the Weak-to-Strong Contrastive Alignment Training Stage, the β parameter in Direct Preference Optimization (DPO) controls the strength of the Kullback-Leibler penalty, typically set within the range of 0.1 to 0.5. A higher β value indicates a preference for the policy model's training process to remain closer to the initially referenced model. In extreme cases, as $\beta \to 0$, we ignore the constraints imposed by the reference model. This setting aims to balance the model's ability to adapt to new training signals while maintaining the stability of the learned behaviors from the reference model.

Subsequently, we trained five variants by adjusting β from 0.1 to 0.5 on the model previously trained with G^3 to explore the impact of the hyperparameter β on attribution quality. We evaluated these variants on the ASQA and ELI5 datasets, and the experimental results are shown in Figure 8.

The experimental results indicate that as β increases, the model's performance on attribution gradually decreases. This observation suggests that the first stage of G^3 might introduce a noticeable inconsistency between grounding and attribution. With higher β values, the model struggles to escape the constraints of inconsistent attributed answers, leading to a reduction in attribution quality as β increases.

E Full Results

We present the comprehensive results of our experiments in Tables 9, 10, and 11. Beyond the evaluation metrics related to Correctness and Citation, we

```
Below is an instruction that describes a task, paired with an input that provides further
context. Write a response that appropriately completes the request.

### Instruction:
Extract the relevant content from the provided documents and then use the extracted content to
guide answer generation and cite the sources properly.
### Input:Question: {Question} Documents: {Documents}
### Response:
```

Table 6: Instruction Format for FRONT on Long-form QA.

```
Below is an instruction that describes a task, paired with an input that provides further
context. Write a response that appropriately completes the request.

### Instruction:
Extract the relevant content from the provided documents and then use the extracted content to
provide a list of accurate answers for the given question. Always cite one and only one document
for each answer. Separate answers by commas.
### Input:Question: {Question} Documents: {Documents}
### Response:
```

Table 7: Instruction Format for FRONT on Short-form QA.

Hyper-parameters	Value
Top-p	0.95
Temperature	1.0
Max-length	2048

Table 8: Hyper-parameter settings in inference.

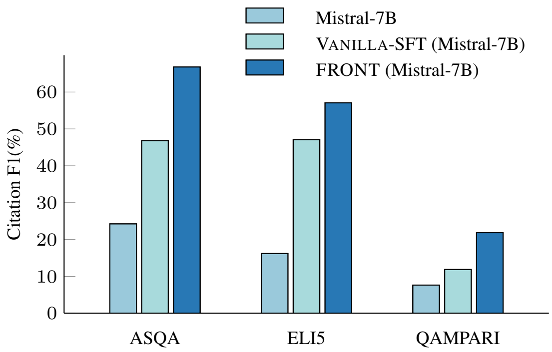

Figure 7: Ablation study on model architecture: We substituted the foundation model in FRONT with Mistral-7B and compared the experimental results of models under the same foundation model using in-context learning and those directly supervised fine-tuned on our automatically generated data. The experiments demonstrate that by replacing different foundation models, our framework still maintains its generalizability.

adhere to the evaluation framework established in (Gao et al., 2023b). For long-form QA datasets like ASQA and ELI5, we also report metrics related to Fluency, ROUGE-L, and average response length. Specifically, we use MAUVE (Pillutla et al., 2021) to evaluate the fluency of the model response. For datasets like QAMPARI, where answers are com-

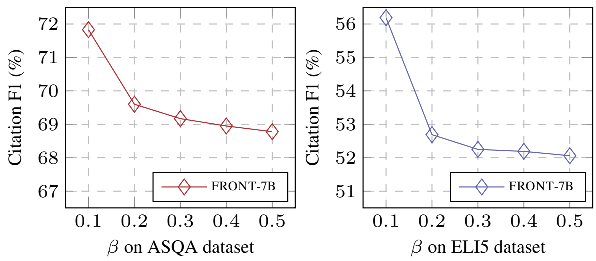

Figure 8: Ablation on hyperparameter β in Weak-to-Strong Contrastive Alignment stage on ASQA and ELI5

posed of concatenated entities, we calculate the average number of predicted entities.

Model Type	**Model Size**	**Fluency** (MAUVE)	**Correct.** (EM Rec.)	**Citation** Rec.	Prec.	F1	ROUGE-L	Length
				Prompting-based				
ChatGPT	-	73.41	40.37	72.81	69.69	71.22	37.92	39.24
LLaMA-2	7B	79.90	24.32	17.24	17.87	17.55	29.38	42.29
	13B	87.08	27.99	16.45	19.04	17.65	31.41	39.25
	70B	69.28	31.53	44.18	44.79	44.48	31.53	26.86
LLaMA-2-Chat	7B	66.78	29.93	55.99	51.66	53.74	32.93	26.18
	13B	66.14	34.39	37.15	38.17	37.65	35.13	33.68
	70B	86.60	41.24	60.19	61.16	60.67	37.01	47.09
Vicuna-v1.5	7B	86.92	38.34	48.37	44.63	46.42	35.95	63.90
	13B	66.11	35.20	51.92	53.40	52.65	35.74	38.57
Mistral	7B	82.37	29.46	23.12	25.45	24.23	31.67	37.17
	8 × 7B	83.30	36.30	32.72	34.49	33.58	35.05	38.47
Mistral-Instruct	7B	82.86	38.57	64.90	59.67	62.18	36.21	45.26
	8 × 7B	94.77	44.11	61.80	63.27	62.53	38.54	58.83
				Post-hoc Retrieval-based				
ChatGPT	-	49.78	37.68	27.11	27.05	27.08	36.64	52.61
LLaMA-2	7B	75.56	16.55	13.88	13.86	13.87	26.81	37.50
	13B	77.91	20.51	20.95	20.94	20.94	29.53	31.37
	70B	75.23	27.58	28.43	28.43	28.43	30.33	29.88
LLaMA-2-Chat	7B	22.50	14.17	11.33	11.33	11.33	21.17	110.04
	13B	64.52	24.43	21.43	21.43	21.43	33.91	41.12
	70B	70.63	29.68	24.51	24.51	24.51	34.17	45.74
Vicuna-v1.5	7B	63.87	19.58	16.24	16.24	16.24	33.22	41.80
	13B	73.83	24.79	24.11	24.11	24.11	34.42	43.54
Mistral	7B	86.54	21.17	16.78	16.77	16.77	30.90	42.43
	8 × 7B	80.99	36.30	38.37	35.27	36.75	35.05	38.47
Mistral-Instruct	7B	67.97	26.26	17.87	17.85	17.86	33.71	51.56
	8 × 7B	65.51	33.90	24.57	24.48	24.52	36.20	53.83
				Training-based				
Self-RAG	7B	74.33	29.96	66.97	67.82	67.39	35.70	29.83
	13B	71.59	31.66	70.35	71.26	70.80	36.01	27.03
VANILLA-SFT	7B	76.66	40.32	67.67	63.67	65.61	38.32	62.00
	13B	84.36	40.85	71.49	66.21	68.75	38.22	58.82
FRONT	7B	81.88	40.84	77.70	69.89	73.59	36.95	53.93
	13B	76.11	41.51	78.44	73.66	75.95	38.63	57.56

Table 9: ASQA full results.

Model Type	Model Size	Fluency (MAUVE)	Correct. (Claim)	Citation Rec.	Citation Prec.	Citation F1	ROUGE-L	Length
Prompting-based								
ChatGPT	-	44.65	12.47	49.44	47.05	48.22	20.64	90.2
LLaMA-2	7B	63.72	4.53	3.92	5.38	4.54	18.27	103.36
	13B	62.19	7.77	8.49	8.43	8.46	19.95	88.23
	70B	53.39	10.43	23.75	22.43	23.07	20.43	93.84
LLaMA-2-Chat	7B	32.80	12.47	19.90	15.48	17.41	20.88	96.42
	13B	29.08	13.83	16.50	16.09	16.29	21.04	94.32
	70B	33.69	13.30	36.63	36.63	36.63	21.29	117.84
Vicuna-v1.5	7B	31.45	12.30	29.81	22.45	25.61	21.36	105.68
	13B	37.41	14.33	31.15	28.99	30.03	21.74	98.23
Mistral	7B	56.62	8.47	16.04	16.32	16.18	20.46	93.80
	8 × 7B	61.83	10.43	26.11	25.09	25.59	20.66	93.59
Mistral-Instruct	7B	32.74	11.07	49.25	42.69	45.74	20.75	98.28
	8 × 7B	38.51	13.93	49.28	48.34	48.81	21.34	113.71
Post-hoc Retrieval-based								
ChatGPT	-	22.79	18.77	14.55	14.55	14.55	22.28	106.83
LLaMA-2	7B	72.80	7.23	6.84	6.84	6.84	19.14	88.19
	13B	53.21	10.33	9.61	9.61	9.61	20.63	90.44
	70B	58.97	11.10	10.27	10.26	10.26	20.41	77.85
LLaMA-2-Chat	7B	22.50	14.17	11.33	11.33	11.33	21.17	110.04
	13B	30.36	14.93	12.10	12.10	12.10	21.82	109.79
	70B	37.87	16.03	12.93	12.93	12.93	21.57	99.94
Vicuna-v1.5	7B	30.88	11.83	10.91	10.91	10.91	21.66	99.03
	13B	32.59	15.20	14.06	14.06	14.05	14.05	108.16
Mistral	7B	52.45	10.47	8.64	8.64	8.64	20.48	90.17
	8 × 7B	48.39	13.57	11.62	11.62	11.62	21.43	91.97
Mistral-Instruct	7B	27.41	17.07	13.20	13.20	13.20	21.52	106.93
	8 × 7B	27.60	17.37	15.68	15.68	15.68	21.66	95.21
Training-based								
Self-RAG	7B	39.14	8.20	8.49	11.80	9.88	17.83	41.70
	13B	37.97	9.20	5.90	8.20	6.86	17.82	43.70
VANILLA-SFT	7B	44.12	9.63	42.30	40.06	41.15	20.58	80.43
	13B	46.33	10.27	46.75	44.47	45.58	20.56	84.01
FRONT	7B	36.90	10.18	58.60	55.33	56.92	19.09	74.06
	13B	34.37	10.32	60.31	59.21	59.75	19.66	75.14

Table 10: ELI5 full results.

Model Type	Model Size	Correctness		Citation			Num Pred.
		Rec.-5	Prec.	Rec.	Prec.	F1	
Prompting-based							
ChatGPT	-	20.28	19.84	19.06	22.03	20.44	4.71
LLaMA-2	7B	12.56	11.32	6.03	6.35	6.19	7.02
	13B	18.00	12.39	5.45	5.74	5.59	11.31
	70B	18.50	14.79	10.10	10.50	10.30	8.31
LLaMA-2-Chat	7B	17.96	19.74	9.58	9.68	9.63	4.73
	13B	21.34	18.86	8.94	9.06	9.00	6.51
	70B	22.62	18.04	13.49	13.98	13.73	7.44
Vicuna-v1.5	7B	14.22	14.74	11.26	11.64	11.45	5.87
	13B	22.06	19.60	13.04	13.74	13.38	7.62
Mistral	7B	16.96	15.98	7.50	7.76	7.63	6.29
	8 × 7B	18.18	15.63	9.72	10.20	9.95	6.63
Mistral-Instruct	7B	17.52	21.29	17.56	18.53	18.03	4.54
	8 × 7B	20.12	19.64	19.27	20.38	19.81	5.32
Post-hoc Retrieval-based							
ChatGPT	-	25.14	22.85	12.29	12.29	12.29	5.46
LLaMA-2	7B	6.48	5.11	5.05	5.05	5.05	6.55
	13B	9.88	7.17	5.20	5.20	5.20	6.98
	70B	14.44	12.44	7.49	7.49	7.49	7.41
LLaMA-2-Chat	7B	12.94	10.89	7.76	7.76	7.76	5.99
	13B	15.72	12.23	7.87	7.87	7.87	6.32
	70B	17.90	14.45	9.05	9.05	9.05	6.05
Vicuna-v1.5	7B	12.04	9.71	6.69	6.69	6.69	7.10
	13B	14.78	11.47	8.50	8.50	8.50	6.67
Mistral	7B	9.94	7.90	6.00	6.00	6.00	7.38
	8 × 7B	13.92	12.08	6.70	6.70	6.70	6.58
Mistral-Instruct	7B	15.80	12.15	8.34	8.34	8.34	7.01
	8 × 7B	24.16	18.28	9.78	9.78	9.78	7.37
Training-based							
VANILLA-SFT	7B	12.86	21.09	21.35	21.36	21.35	7.49
	13B	12.68	22.80	23.64	23.71	23.67	3.14
FRONT	7B	11.50	21.38	24.74	24.84	24.79	3.08
	13B	11.94	22.61	24.86	25.39	25.12	3.17

Table 11: QAMPARI full results.