Transparentize the Internal and External Knowledge Utilization in LLMs with Trustworthy Citation

Jiajun Shen[1,3*], **Tong Zhou**[1*], **Yubo Chen**[1,2†],
Delai Qiu[4], **Shengping Liu**[4], **Kang Liu**[1,2,5], **Jun Zhao**[1,2†]
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]Unisound AI Technology Co,Ltd [5]Shanghai Artificial Intelligence Laboratory
shenjiajun21@mails.ucas.ac.cn, tong.zhou@ia.ac.cn, {yubo.chen,jzhao}@nlpr.ia.ac.cn

Abstract

While hallucinations of large language models could been alleviated through retrieval-augmented generation and citation generation, how the model utilizes internal knowledge is still opaque, and the trustworthiness of its generated answers remains questionable. In this work, we introduce Context-Prior Augmented Citation Generation task, requiring models to generate citations considering both external and internal knowledge while providing trustworthy references, with 5 evaluation metrics focusing on 3 aspects: answer helpfulness, citation faithfulness, and trustworthiness. We introduce RAEL, the paradigm for our task, and also design INTRALIGN, an integrated method containing customary data generation and an alignment algorithm. Our experimental results show that our method achieves a better cross-scenario performance with regard to other baselines. Our extended experiments further reveal that retrieval quality, question types, and model knowledge have considerable influence on the trustworthiness in citation generation.

1 Introduction

Large Language Models (LLMs; Brown et al., 2020) have demonstrated remarkable question-answering (QA) capabilities, providing users with helpful response (Shaier et al., 2024). However, due to the hallucination of LLMs, it is crucial to improve the trustworthiness of the responses (Liu et al., 2023; Zhou et al., 2024). Retrieval Augmented Generation (RAG) with explicit citations can utilize the retrieved external knowledge and link the response to the knowledge to improve the transparency of LLM's response and increase user trust (Ding et al., 2025). However, previous work (Appendix A) on leveraging context and prior knowledge in LLM generation pays minor attention to the following two problems: (1) the interpretability of LLMs when utilizing prior knowledge and (2) the trustworthiness of the reference.

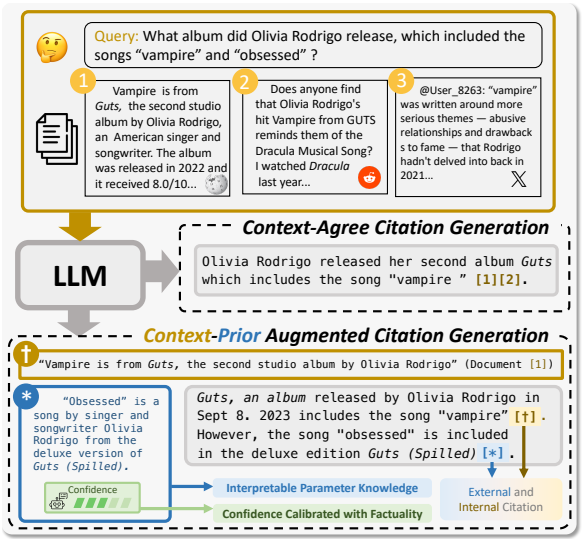

Figure 1: Compared with Context-Agree Citation Generation, the Context-Prior Augmented Citation Generation allows LLMs to appropriately utilize and cite parameter knowledge in an interpretable way, and requires LLMs to extract convincing and concise external references, aiming at transparentize the internal and external knowledge utilization as well as enhancing trustworthiness.

Interpretability of prior knowledge utilization. Prior knowledge, or parameter knowledge, referring to the knowledge encoded in the model's parameters, could serve as a supplement when the retriever fails to meet the needs of the question in RAG (Sun et al., 2023). For example, as indicated in Figure 1, the question *"What album did Olivia Rodrigo release, which included the songs 'vampire' and 'obsessed'?"* contains two constraints, but the external documents only provide clues about one of them. If the model has supplementary prior knowledge, it can articulate the knowledge, thereby making the answer accurate and providing verifiable evidence. Despite the significance of LLM's prior knowledge, previous studies (Yu et al., 2024;

[*]These authors contributed equally to this work.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 17858–17877

Sun et al., 2023; Minder et al., 2024; Ming et al., 2024; Cheng et al., 2024) have largely overlooked its interpretable utilization in citation generation tasks. Therefore, appropriately articulating and citing reliable parameter knowledge in an interpretable way remains challenging.

Trustworthiness in reference. Citations are important in enhancing the convincingness and verifiability of AI-generated content. (Ding et al., 2025). Highly convincing cited references should contain complete and self-consistent information to support the answer. Concise references reduce the user's verification cost, increasing their trust in the system. For example, Figure 3 shows two contrastive cases. The reference `[1]` is concise but lacks background information and supporting evidence, making it doubtful, though easy to verify. The reference `[6]` provides background and details, making it convincing, but includes abundant distracting background information about the question, such as the person's personality, making it less pithy and increasing the cost of verification. The trustworthiness impacts a user's acceptance and trust in the reference, but previous work has not considered this. Furthermore, since there are constraints between convincingness and conciseness, improving both aspects is challenging.

To fill the gap in the interpretability of prior knowledge in citation tasks, we propose Context-Prior Augmented Citation Generation task, which requires the model to generate and cite references from prior knowledge if needed to improve the quality of the response and report a confidence score to transparentize the utilization of prior knowledge. To conduct comprehensive evaluations, we design 5 metrics: (1) Accuracy for the helpfulness of the answer, (2) Citation Recall for citation faithfulness, (3) Convincingness, (4) Conciseness, and (5) Expected Calibration Error for the trustworthiness of the reference. Our evaluations demonstrate a strong correlation between automatic metrics and human judgments.

In response to the requirement of our task, we propose RAEL (**R**ational **A**ttribution and **El**aboration), a paradigm to enable LLMs to use internal knowledge and generate trustworthy citations appropriately. For the interpretability of parameter knowledge and the faithfulness of citations, we design INTRALIGN (**In**terpretable **Tr**ustworthiness **Align**ment) to obtain dataset by incorporating the parameter knowledge of the LLM and an alignment step, which enables the LLM to generate faithful and trustworthy citations.

We conduct experiments with different LLM citation generation methods. Our experiments successfully reveal that existing citation generation methods struggle to adapt to scenarios with poor retrieval performance and to cite trustworthy references. INTRALIGN leads to considerable improvements across all metrics, demonstrating strong practicality. Our contributions can be summarized as follows:

- We propose Context-Prior Augmented Citation Generation task requiring models to appropriately generate and cite references from prior knowledge and design 5 complementary metrics to evaluate helpfulness, faithfulness, and trustworthiness of LLM's response.
- We introduce RAEL as the paradigm for our task and INTRALIGN, which contains multi-scenario trustworthy data generation and interpretability-focused alignment, allowing the model to utilize prior knowledge and generate trustworthy responses.
- We evaluate 6 baselines and our proposed method with 3 LLMs across 4 scenarios. Experiments reveal the shortcomings of existing methods in improving the overall performance on our task, and our method enables the model to cite references from parameter knowledge while effectively improving the quality of references and enhancing their trustworthiness.

2 Task and Metrics

In this section, we present a formal definition of the general citation generation task and give the definition of Context-Prior Augmented Citation Generation task along with the metrics introduced.

Context-Agree Citation Generation. The citation generation task accepts a question along with context sequence D and returns an answer, which can be split into t segments $(S_1, S_2, \ldots, S_t)$. Each segment S_i is paired with a reference R_i, and we define $R_i = \mathcal{F}(S_i)$ as segment S_i cites reference R_i, or S_i has no citation if $R_i = \varnothing$. Segment S_i is usually split by sentence boundaries (Gao et al., 2023b; Huang et al., 2024b), and the paired reference is a full document in coarse-grained citations. In fine-grained citation generation (Xu et al., 2024b; Zhang et al., 2024b), the reference can span fewer words. To ensure the citation's faithfulness to the context, the cited text must be verbatim: R_i should be a subsequence from D. When a question

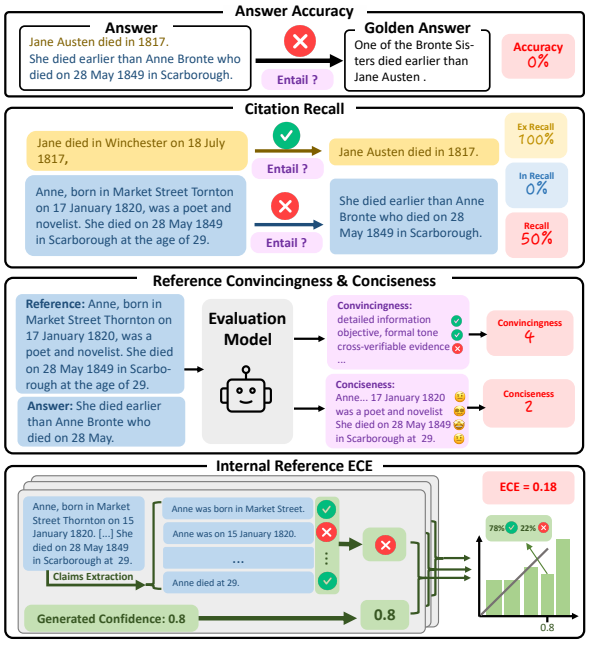

Figure 2: Illustration of our metrics and the auto evaluation process. We use the same NLI model to check entailment to prevent bias.

is unanswerable based on D, LLM should generate a refusal answer to stay faithfulness to contexts.

Context-Prior Augmented Citation Generation. Though faithfulness to the context reduces hallucination, parameter knowledge of LLMs can be beneficial in circumstances of insufficient external sources or low-quality retrieval. To enable LLM to cite parameter knowledge, we require the reference to be from the context or prior knowledge. In our definition, reference $\mathcal{F}(S_i)$ should be either (1) a non-empty extraction from D_i as external reference $R_i^{ex} \neq \varnothing$, or (2) a sequence $(R_i^{in}; P_i)$ as internal reference, where (;) denotes concatenation and P_i denotes a confidence score of $R_i^{in} \neq \varnothing$. The confidence score represents the estimation of the factuality of the reference. A refusal answer is preferred if and only if the question is unanswerable based on D and the LLM's parameter knowledge.

2.1 Metrics

To implement a comprehensive evaluation in Context-Prior Augmented Citation Generation, we measure **Accuracy** for the helpfulness of the answer, Citation **Recall** for the faithfulness of citations, Reference **Convincingness**, **Conciseness**, and **Expected Calibration Error** for the trustworthiness of the reference. We will introduce these metrics below and explain how the metrics ensure robustness against shortcuts in Appendix C.

2.1.1 Answer Accuracy

Accuracy measures how the response generated by LLM correctly answers the input question. We use a semantics-based method, using an NLI model to verify whether the model's response entails the golden sentence. The NLI model returns a bool value $\phi(a, g) = 1$ if the answer a entails the golden answer g. The average accuracy on the dataset is calculated as $\frac{1}{N}\sum_{i=1}^{N}\phi(a_i, g_i)$, where N is the size of the dataset and a_i, g_i are the answer and golden answer from i-th sample. We illustrate that our method reduces FN and FP by 25.14% and 44.93%, compared to String Exact Match (Stelmakh et al., 2022) in §6, respectively.

2.1.2 Citation Recall

Recall shows how faithful the response is to the original reference. Following Gao et al.'s (2023b) work, we also use an NLI model to verify whether the cited reference entails the response. We calculate Rc^O as the overall Recall. Since we observe different recall scores in external citations and internal citations, we also divide S_i where $\mathcal{F}(S_i) \neq \varnothing$ into two sets S^{ex} and S^{in} according to the cited reference, and calculate each type independently as Rc^{in}, Rc^{ex}. We exclude refusal answers when calculating recall scores. Formally, given a citation function $\mathcal{F}$ and statements S, the average Recall scores Rc^O is $Rc^O = \frac{1}{|S|}\sum_{S_i \in S}\phi(\mathcal{F}(S_i), S_i)$.

We only use $S_i^{ex} \in S^{ex}$ and $S_i^{in} \in S^{in}$ to calculate Rc^{in} and Rc^{ex}, respectively. Sentences without citations ($\mathcal{F}(S_i) = \varnothing$) is not included in the computation of Rc^{in} and Rc^{ex} but will lower Rc^O.

2.1.3 Reference Convincingness

Convincingness measures how the cited reference is trustworthy to humans. We expect LLMs to cite convincing references with formal and objective language style, complete expressions, unambiguous entity references, and coherent logic. Fully relying on humans to evaluate this subjective metric is time-consuming, so we use a strong LLM aligned with human preference as the evaluation model for automatic evaluation. When evaluating, we mask all the entities in the reference to avoid bias from prior knowledge and then ask the evaluation model to generate a score from 1 to 5 with an explanation considering the following aspects: tone, style, objectivity, logical coherence, disambiguation, and richness of evidence. We show our prompt in Appendix G. We also conduct experiments in §6 to verify human and automatic evaluation alignment.

2.1.4 Reference Conciseness

Conciseness reflects the subjective cost required by a person when verifying information. Excessive distracting content, such as too much background information, can lead to wasted time in reading and verification. Conciseness is not the average relevance of sentences to the answer because appropriate background information is helpful. We use an evaluation model to simulate the human process of reading sentence-by-sentence, assessing in sequence whether the text provides useful information with minimal distractions, and report a score from 1 to 5. We show our prompt in Appendix G. The trade-off between Convincingness and Conciseness, as shown in Figure 3, makes it challenging to improve both metrics simultaneously, even if the reference is already objective and coherent.

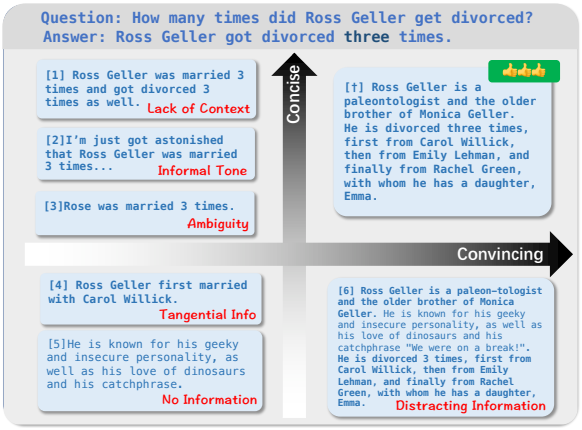

Figure 3: Example of different Convincingness and Conciseness Scores

2.1.5 Internal Reference ECE

In our task, we ask LLMs to report a confidence score when generating a reference from parameter knowledge. To measure the faithfulness of the confidence score generation, we use Expected Calibration Error (ECE) to measure the alignment between the confidence of the output reference P and its real factuality. We evaluate the correctness of each internal reference R_i^{in} using FACTSCORE (Min et al., 2023) and assume the reference is correct if all the facts are correct (i.e., $\text{FS}(R_i^{in}) = 1$, FS returns the factuality of a reference). We assign each i in the index set to m bins, $B_1, B_2, \ldots, B_m$, based on P_i, where for any $j \in B_k$, $P_j \in \left(\frac{k-1}{m}, \frac{k}{m}\right]$, and then the ECE is calculated as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \cdot |\text{fact}(B_m) - P(B_m)|$$

where $\text{fact}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{I}(\text{FS}(R_i^{in}) = 1)$ and $P(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} P_i$.

3 Dataset

We construct the dataset from the three latest RAG datasets: CRAG(Yang et al., 2024), FRAMES(Krishna et al., 2024), and SituatedFaithfulnessEval(Huang et al., 2024c) (SFE), as they provide diverse, challenging questions.

We combine the three datasets, equip each data point with 5 retrieved documents, and annotate whether the document is ground truth. (i.e., contains the answer). According to whether the data point contains a **G**round **T**ruth document or not, we split the dataset into 2 settings **GT** and $\overline{\textbf{GT}}$, respectively. Detailed dataset profile and annotation step are shown in Appendix D.

4 Method

Based on the proposed task requirements, we designed a generation paradigm and an integrated method for aligning open-source models. We propose a **Rational Attribution and Elaboration** (RAEL) paradigm to align models with the requirements of our proposed task. Specifically, we asked to review the context and scrutinize parameter knowledge to help selectively use context and faithfully state parameter knowledge. Then, the model extracts context and recites parameter knowledge to provide trustworthy references. We design INTRALIGN (**In**terpretability-**Tr**ustworthiness **Align**ment, a pipeline using reject sampling to generate customary data from GPT-4o (OpenAI, 2024) and use the data tailored for the specific target model to enhance its performance, as in Figure 4.

4.1 Rational Attribution and Elaboration

To ensure the model stays faithful to both the internal and external information in generating, we require the model to review the context, scrutinize its own knowledge, and then generate a series of context excerpts or self-generated parameter knowledge as references, along with a final answer containing citations. Reviewing and scrutiny enable the model to rationally consider the sufficiency of external information and the potential relevant knowledge from within, enhancing faithfulness and explainability. The example is shown in Figure 4, with a Context Review, Parameter Knowledge, References (Including extracted context and recited

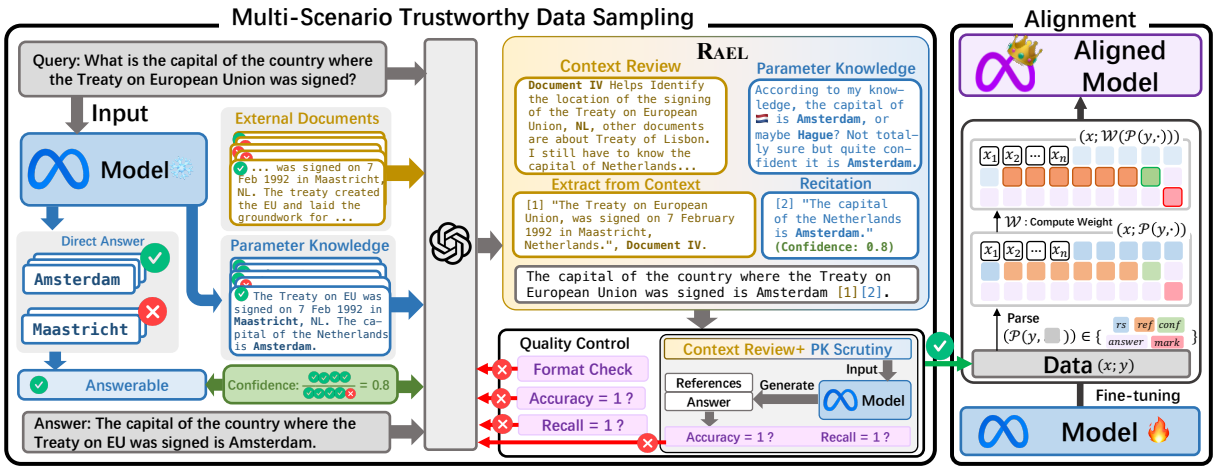

Figure 4: Overview of INTRALIGN. We first conduct multi-scenario trustworthy data sampling to incorporate parameter knowledge and generate a golden response following our RAEL paradigm. The verified high-quality data will be used for subsequent Interpretability-Focused Alignment, ultimately resulting in a model capable of utilizing parameter knowledge and generating trustworthy citations.

internal knowledge), and an Answer.

4.2 Interpretable Trustworthiness Alignment

To improve the model's faithfulness and trustworthiness for external and internal knowledge in citation generation, we propose a pipeline that samples k documents and k direct answers from the target LLM. If any documents are ground truth or any answer is correct, the LLM is regarded with the necessary parameter knowledge to the question. According to whether the LLM has corresponding **P**arameter **K**nowledge, we divide the dataset into two categories **PK** and $\overline{\textbf{PK}}$, respectively.

For all sampled documents $d_1, d_2, ..., d_k$ for each question, we use an NLI model $\phi(\cdot)$ to check whether they contain the answer g. The golden confidence is $\frac{\sum_1^k \phi(d_i, g)}{k}$ (The possibility of consistently generating a document with the golden answer). The formula is inspired by self-consistency-based uncertainty measurement (Wang et al., 2023) in LLMs, but we additionally use the NLI model to ensure factuality.

Given the external documents and internal documents, we asked GPT-4o to generate multiple responses with the RAEL paradigm using the prompt in Figure 14. We evaluate the generated results and regenerate data with incorrect answers or unfaithful citations. Next, we use only the questions, the context review, and parameter knowledge scrutiny to make the target model generate answers and references. We regenerate responses with incorrect answers or unfaithful citations to ensure the quality of the review and scrutiny step. Finally, we rerank the responses by Convincingness and Conciseness and select the one with the highest score.

Due to dispersed optimization objectives in alignment and the neglect caused by limited token usage for citation markers and confidence scores, we adjusted the weights of different types of tokens during the alignment step to make the model focus more on interpretable and trustworthy information. We compute token-wise weighted loss to ensure focus on citation generation. We parse label token sequence in the dataset, and for each token y_t in the label, the parser function $\mathcal{P}$ maps the label sequence y and index t to a type $\tau = \mathcal{P}(y, t) \in \mathcal{T}$, where $\mathcal{T} = \{\tau_{rs}, \tau_{ref}, \tau_{answer}, \tau_{conf}, \tau_{mark}\}$, each denoting review and scrutiny, reference, answer, confidence, and citation markers (such as `[1][2]`). Each type $\tau \in \mathcal{T}$ is assigned with a weight $w = \mathcal{W}(\tau)$. The loss of the i-th output $y^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_{T_i}^{(i)})$ given the input $x^{(i)}$ is

$$\mathcal{L} = \sum_{t=1}^{T_i} \mathcal{W}(\mathcal{P}(y^{(i)}, t)) \log P_\theta(y_t^{(i)} | y_{<t}^{(i)}, x^{(i)})$$

As for the weights, to ensure that the loss gives sufficient attention to confidence, reference and citation markers, we define the total weights for type τ as $\hat{\mathcal{W}}(\tau) = \sum_{i=0}^{T_i} \mathcal{W}(\tau) \cdot \mathbb{I}(\mathcal{P}(y, t) = \tau)$, and set $\hat{\mathcal{W}}(\tau_{conf}) = \hat{\mathcal{W}}(\tau_{ref}), \hat{\mathcal{W}}(\tau_{mark}) = \hat{\mathcal{W}}(\tau_{answer})$. Since our metrics highly depend on the references, we fix the weights of review, scrutiny, and answer to $\mathcal{W}(\tau_{rs}) = \mathcal{W}(\tau_{answer}) = 1$ and increase the weight of reference such that $\hat{\mathcal{W}}(\tau_{ref}) = \hat{\mathcal{W}}(\tau_{rs}) + \hat{\mathcal{W}}(\tau_{answer})$ to make the model focus

more on reference generation. $\mathcal{W}$ can be determined given the constraints above.

As indicated in Figure 4, after having determined the weights, we use function $(x; \mathcal{P}(y, \cdot))$ to convert the tokenized input into a tensor of labels, and then apply $(x; \mathcal{W}(\mathcal{P}(y, \cdot))$ to convert the labels to different weights. We use darkness to represent the weight in the figure (e.g., the orange tokens are darker since they represent the references).

This interpretability-focused alignment with dynamic weights improves the model's trustworthiness and interpretability without sacrificing its overall performance, as shown in the results in §5.5.

5 Experiments

We conduct comprehensive experiments on three LLMs with baselines and our method. Then, we present the results along with a detailed and thorough analysis.

5.1 Settings

To showcase the influence of parameter knowledge and instruction following ability, we use two open-source models from the same family with two parameter sizes: Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct (AI@Meta, 2024). We also apply a more powerful closed-source LLM, GPT-4o from OpenAI (OpenAI, 2024). As deep thinking CoTs in inference like OpenAI-o1 and DeepSeek-R1 models are popular, and their "thinking" part contains the description about how they use external knowledge and internal knowledge, we also use o1-mini and DeepSeek-R1 as our models. We ask the LLM to provide an answer in the RAEL framework, without giving examples for the reasoning step, to make sure the reference model can freely choose their own thinking style.

We obtain 1K training and 0.5K test data for training and evaluation, using GPT-4o-mini to evaluate Convincingness and Conciseness.

5.2 Baselines

We compare our method with the state-of-the-art methods on RAG and citation generation. We use 6 baselines in our experiments, including: (1) Guided-RAEL uses a two-shot prompt to guide the model in applying RAEL paradigm. (2) FOOTNOTE generates answers with reference in the format `\footnote[confidence]{reference}`. (3) POSTCITE retrieves documents using a GTR retriever and cites the document with the highest similarity score unless it falls below a threshold, in which case the model generates an internal citation. (4) RECITATION AUGMENTED GENERATION samples passages from the model's parameters and generates answers based on them, determining the final answer via majority voting. (5) FRONT optimizes citation quality by extracting supporting quotes and ensuring consistency. (6) CONTEXTCITE identifies the specific parts of the context that contribute to the response using sparse linear modeling. Details about baselines are shown in E.1.

5.3 Main Results

We show our main results on the **GT**, **PK**; **GT**, $\overline{\textbf{PK}}$ and $\overline{\textbf{GT}}$, **PK** sets in Table 1 and detail the findings below. We also provide further analysis of refusal answers on the $\overline{\textbf{GT}}$, $\overline{\textbf{PK}}$ set in Appendix F.

A notable difference lies between $\textbf{Rc}^{\textbf{ex}}$ and $\textbf{Rc}^{\textbf{in}}$. For methods utilizing both external and internal citation except FOOTNOTE, the internal citation recall Rc^{in} is generally higher than the external citation recall Rc^{ex}. This indicates that models are more faithful in citation when using internal knowledge, and the `\footnote` format is difficult for LLMs to follow when citing prior knowledge.

Existing methods struggle with cross-scenario performance and trustworthiness. FRONT and CONTEXTCITE suffer from a nearly 30% drop in accuracy and Rc^{ex} when the retrieval quality is low. RECITE, FRONT, and CONTEXTCITE also have shortcomings in terms of trustworthiness, exhibiting lower Convincingness, Conciseness, and higher ECE.

Our method achieves better overall performance with more trustworthy references. Our approach achieves outstanding performance across all major metrics in each scenario. Notably, in terms of citation quality, our method minimizes the proportion of uncited sentences as much as possible while maintaining high recall scores for both external and internal citations with Rc^{in} higher than 80%. Additionally, the references generated by our method exhibit better trustworthiness, with a score of nearly or higher than 4.00 in terms of the Conciseness score. The alignment between the model's confidence in parameter knowledge and factual accuracy in our method surpasses most baselines, indicating that the references produced by our model are of higher quality and more trustworthy.

LLMs can learn to use external and internal knowledge adaptively. Our method maintains high accuracy across different scenarios. In the **GT**, $\overline{\textbf{PK}}$ and $\overline{\textbf{GT}}$, **PK** sets, the model achieves sim-

Scenario	Model	Method	Helpfulness	Recall ↑			Trustworthiness		
			Accuracy ↑	Rc^{ex}	Rc^{in}	Rc^{O}	Conv. ↑	Conc. ↑	ECE ↓
GT, PK	Llama-3.1-8B-Instruct	RECITE†	61.38 (2.78)	-	65.71 (1.02)	60.23 (1.04)	3.37 (0.03)	3.36 (0.05)	0.29 (0.04)
		FRONT‡	72.08 (1.49)	**71.84** (1.84)	-	48.00 (1.50)	3.42 (0.28)	2.01 (0.07)	-
		CONTEXTCITE‡	71.10 (4.95)	25.25 (1.43)	-	25.08 (1.29)	3.43 (0.01)	3.47 (0.01)	-
		FOOTNOTE	65.91 (2.09)	51.20 (1.65)	44.38 (1.81)	16.03 (1.63)	3.32 (0.1)	3.96 (0.12)	0.20 (0.02)
		POSTCITE	71.64 (3.88)	31.01 (3.04)	30.77 (2.58)	29.82 (1.84)	3.47 (0.08)	2.25 (0.12)	0.14 (0.02)
		Guided-RAEL	62.87 (4.43)	61.54 (2.06)	74.46 (1.73)	57.33 (1.49)	3.14 (0.09)	3.97 (0.05)	0.12 (0.03)
		INTRALIGN (Ours)	**75.90** (2.80)	67.63 (1.46)	**85.80** (1.51)	**63.72** (1.02)	**3.61** (0.11)	**4.05** (0.07)	**0.10** (0.02)
	Llama-3.1-70B-Instruct	RECITE†	45.05 (2.46)	-	81.62 (5.15)	77.86 (3.34)	3.54 (0.15)	3.86 (0.06)	0.22 (0.03)
		FRONT‡	76.65 (1.53)	**72.82** (6.34)	-	57.61 (3.90)	3.26 (0.07)	2.53 (0.06)	-
		CONTEXTCITE‡	72.49 (1.34)	33.08 (3.92)	-	33.06 (2.23)	3.51 (0.03)	3.19 (0.09)	-
		FOOTNOTE	75.83 (4.30)	52.91 (3.65)	32.14 (1.44)	30.36 (1.12)	3.43 (0.14)	4.10 (0.04)	0.23 (0.00)
		POSTCITE	64.82 (2.09)	23.86 (2.32)	66.29 (2.13)	43.71 (1.76)	3.69 (0.11)	1.81 (0.10)	0.19 (0.02)
		Guided-RAEL	73.06 (6.20)	62.00 (4.93)	66.67 (2.11)	59.79 (1.19)	3.32 (0.04)	3.97 (0.03)	0.13 (0.01)
		INTRALIGN (Ours)	**85.72** (3.13)	68.45 (1.20)	**88.10** (1.01)	**78.79** (1.27)	**3.69** (0.19)	**4.42** (0.14)	**0.10** (0.01)
	GPT-4o	POSTCITE	81.81 (6.18)	39.54 (3.17)	**81.95** (2.46)	56.32 (1.79)	2.94 (0.14)	2.95 (0.04)	0.20 (0.02)
		FOOTNOTE	**82.19** (5.45)	57.22 (6.85)	52.14 (4.27)	51.75 (5.60)	3.42 (0.04)	3.61 (0.16)	0.18 (0.02)
		Guided-RAEL	81.59 (10.76)	**62.94** (7.31)	66.67 (6.62)	**58.84** (7.01)	**3.58** (0.07)	**4.00** (0.14)	0.10 (0.00)
	DeepSeek-R1	FOOTNOTE	90.55 (–)	56.78 (–)	50.33 (–)	51.25 (–)	3.20 (–)	3.87 (–)	0.17 (–)
		GUIDED-RAEL	84.77 (–)	53.89 (–)	49.47 (–)	51.45 (–)	3.28 (–)	4.04 (–)	0.15 (–)
	o1-mini	FOOTNOTE	77.53 (–)	53.92 (–)	43.56 (–)	49.01 (–)	3.26 (–)	3.89 (–)	0.15 (–)
		GUIDED-RAEL	70.88 (–)	56.25 (–)	47.92 (–)	52.17 (–)	3.23 (–)	4.00 (–)	0.13 (–)
GT, $\overline{\text{PK}}$	Llama-3.1-8B-Instruct	RECITE†	0.95 (1.41)	-	56.51 (1.50)	48.49 (1.88)	3.05 (0.03)	2.04 (0.03)	0.23 (0.04)
		FRONT‡	64.29 (1.66)	**68.62** (2.82)	-	56.25 (1.88)	1.80 (0.05)	1.86 (0.05)	-
		CONTEXTCITE‡	56.67 (3.49)	35.56 (1.14)	-	35.55 (1.20)	3.51 (0.16)	3.01 (0.03)	-
		FOOTNOTE	50.95 (3.81)	56.52 (1.90)	45.90 (1.19)	14.16 (1.23)	3.55 (0.02)	3.47 (0.07)	0.17 (0.02)
		POSTCITE	63.33 (2.51)	22.03 (1.08)	35.93 (2.03)	24.24 (1.59)	3.60 (0.01)	2.38 (0.20)	0.13 (0.01)
		Guided-RAEL	48.10 (2.45)	51.61 (1.77)	43.75 (1.42)	43.56 (1.39)	3.48 (0.10)	**3.80** (0.08)	0.13 (0.03)
		INTRALIGN (Ours)	**69.05** (2.65)	44.53 (1.43)	**87.95** (1.72)	**51.84** (1.01)	**3.64** (0.03)	3.78 (0.12)	**0.11** (0.02)
	Llama-3.1-70B-Instruct	RECITE†	1.95 (0.25)	-	80.18 (6.52)	78.73 (6.17)	**3.82** (0.1)	3.09 (0.07)	0.23 (0.02)
		FRONT‡	65.24 (0.55)	**72.10** (0.69)	-	53.78 (5.35)	3.48 (0.11)	2.32 (0.07)	-
		CONTEXTCITE‡	63.07 (3.17)	38.89 (5.90)	-	38.86 (0.59)	3.32 (0.13)	3.29 (0.04)	-
		FOOTNOTE	67.74 (6.48)	68.18 (1.97)	33.33 (3.06)	30.92 (1.32)	3.71 (0.12)	4.40 (0.13)	0.18 (0.01)
		POSTCITE	58.13 (1.04)	21.93 (0.63)	68.90 (6.98)	42.56 (2.21)	3.86 (0.10)	1.37 (0.02)	0.15 (0.01)
		Guided-RAEL	63.40 (0.79)	40.44 (4.71)	50.00 (0.26)	40.61 (7.31)	3.54 (0.06)	3.18 (0.02)	**0.10** (0.00)
		INTRALIGN (Ours)	**71.25** (1.44)	54.42 (2.60)	**82.07** (3.76)	**60.32** (4.79)	3.60 (0.08)	**4.47** (0.03)	0.13 (0.03)
	GPT-4o	POSTCITE	**75.00** (2.87)	24.44 (1.03)	**80.67** (7.68)	**52.03** (2.99)	**3.78** (0.12)	2.67 (0.08)	0.18 (0.01)
		FOOTNOTE	66.03 (6.02)	**47.88** (0.03)	52.87 (2.13)	32.38 (2.8)	3.55 (0.00)	3.85 (0.09)	0.20 (0.03)
		Guided-RAEL	69.51 (5.2)	41.00 (3.45)	83.33 (1.78)	42.94 (4.43)	3.47 (0.11)	**4.09** (0.05)	0.18 (0.02)
	DeepSeek-R1	FOOTNOTE	76.02 (–)	41.66 (–)	40.78 (–)	40.96 (–)	3.35 (–)	3.82 (–)	0.18 (–)
		GUIDED-RAEL	76.64 (–)	43.05 (–)	57.14 (–)	41.89 (–)	3.26 (–)	4.00 (–)	0.18 (–)
	o1-mini	FOOTNOTE	64.61 (–)	47.61 (–)	43.06 (–)	45.51 (–)	3.60 (–)	4.05 (–)	0.15 (–)
		GUIDED-RAEL	51.58 (–)	57.00 (–)	40.23 (–)	38.53 (–)	3.53 (–)	4.51 (–)	0.15 (–)
$\overline{\text{GT}}$, PK	Llama3.1-8B-Instruct	RECITE†	61.49 (1.68)	-	66.13 (1.17)	59.81 (2.32)	3.32 (0.02)	3.32 (0.03)	0.30 (0.03)
		FRONT‡	48.60 (3.07)	54.61 (1.28)	-	31.60 (1.56)	3.30 (0.08)	1.86 (0.01)	-
		CONTEXTCITE‡	44.41 (3.99)	12.46 (1.47)	-	12.46 (1.41)	3.37 (0.16)	3.81 (0.15)	-
		FOOTNOTE	35.44 (1.30)	35.46 (1.19)	31.15 (2.11)	18.87 (1.33)	3.59 (0.11)	3.29 (0.08)	0.17 (0.01)
		POSTCITE	48.60 (3.98)	24.70 (2.43)	33.84 (2.47)	28.59 (2.26)	3.62 (0.11)	2.51 (0.12)	0.20 (0.01)
		Guided-RAEL	49.30 (2.97)	42.77 (3.42)	72.34 (2.05)	47.54 (1.82)	3.05 (0.15)	3.59 (0.10)	0.13 (0.03)
		INTRALIGN (Ours)	**62.24** (3.84)	**59.11** (1.25)	**72.65** (2.73)	**59.80** (1.50)	**3.65** (0.13)	**3.83** (0.07)	**0.11** (0.02)
	Llama3.1-70B-Instruct	RECITE†	44.43 (2.01)	-	82.12 (6.18)	72.86 (5.04)	3.44 (0.26)	3.72 (0.19)	0.22 (0.02)
		FRONT‡	56.06 (0.18)	47.46 (0.52)	-	42.28 (3.96)	3.32 (0.02)	2.24 (0.04)	-
		CONTEXTCITE‡	52.91 (1.65)	15.81 (2.19)	-	15.81 (1.62)	3.55 (0.02)	3.86 (0.11)	-
		FOOTNOTE	54.66 (2.08)	41.98 (0.58)	26.23 (0.77)	20.62 (0.07)	3.53 (0.11)	4.10 (0.06)	0.18 (0.01)
		POSTCITE	53.88 (1.64)	22.63 (3.17)	74.21 (3.48)	45.13 (1.34)	3.66 (0.04)	1.74 (0.05)	0.19 (0.01)
		Guided-RAEL	57.32 (7.12)	46.15 (1.25)	81.94 (7.02)	51.64 (4.48)	3.33 (0.09)	3.18 (0.06)	0.12 (0.00)
		INTRALIGN (Ours)	**75.64** (0.72)	**62.01** (4.20)	**89.82** (1.35)	**75.71** (2.19)	**3.67** (0.07)	**4.42** (0.09)	**0.09** (0.00)
	GPT-4o	POSTCITE	**78.59** (5.84)	39.83 (0.81)	**77.26** (1.35)	53.72 (1.65)	3.53 (0.13)	2.94 (0.04)	0.15 (0.00)
		FOOTNOTE	72.89 (6.65)	40.06 (2.91)	53.78 (5.91)	49.51 (8.93)	3.42 (0.15)	4.00 (0.03)	0.15 (0.00)
		Guided-RAEL	72.27 (3.13)	**43.97** (0.59)	69.61 (0.82)	**54.93** (6.22)	**3.64** (0.08)	**4.04** (0.07)	0.22 (0.03)
	DeepSeek-R1	FOOTNOTE	71.66 (–)	36.41 (–)	50.82 (–)	40.40 (–)	3.30 (–)	3.86 (–)	0.18 (–)
		GUIDED-RAEL	73.29 (–)	36.89 (–)	50.32 (–)	42.31 (–)	3.20 (–)	3.94 (–)	0.13 (–)
	o1-mini	FOOTNOTE	61.23 (–)	39.82 (–)	42.50 (–)	40.83 (–)	3.01 (–)	3.81 (–)	0.13 (–)
		GUIDED-RAEL	63.23 (–)	69.23 (–)	51.02 (–)	45.35 (–)	3.09 (–)	4.07 (–)	0.13 (–)

Table 1: Results on test sets **GT**, **PK**; **GT**, $\overline{\textbf{PK}}$ and $\overline{\textbf{GT}}$, **PK**. We use different random seeds to run three experiments for each setting (except for o1-like models) and show the mean scores. The values in brackets represent the standard deviation. Methods marked with † are limited to citing parameter knowledge, while sections marked with ‡ are limited to citing external knowledge. **bold** values represent model-wise best score, and background represents the best score (before rounded to 2 decimals) across models.

ilarly high accuracy. This demonstrates that our method enables the model to leverage both external and internal knowledge adaptively.

It is noticeable that the inference model achieves relatively higher accuracy but fails to provide high-quality citations, especially for internal citations. We thank the reviewer for reminding us of the o1-like model's ability to give a thoughtful and accurate answer. However, the claim that the general CoT process does not necessarily improve citation quality and trustworthiness still holds. Moreover, given the fact that the R1 model outputs 5 times

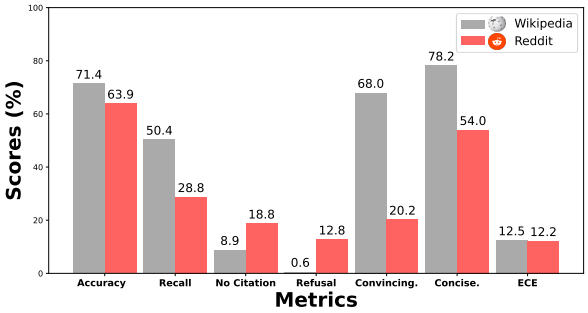

Figure 5: Results on Wikipedia and Reddit dataset. We rescaled each metric to a 0%-100% range.

more tokens than our model (see the table below) due to overthinking, the deep thinking model's performance on our task is still limited.

5.4 In-depth Results of the Task

In this section, we discuss more experimental results on different settings that reveal the characteristics of models on the Context-Prior Augmented Citation Generation task.

5.4.1 Results on a less credible external source

We use GPT-4o to imitate the style of Reddit posts and substitute the documents to generate a set with less convincing documents. We fix other settings and use this dataset to evaluate the model's behavior under the situation with a less convincing knowledge source. We show the results of our experiment on Llama-3.1-8B-Instruct in Figure 5.

Our results show that when external documents are less convincing, the model suffers a significant drop in performance, especially on recall, and is more likely to generate a refusal answer even if a ground truth document is provided. This indicates that the quality of external knowledge sources significantly impacts citation generation tasks.

5.4.2 Tug-of-war between knowledge

The tug-of-war between external and prior knowledge in RAG, especially in conflict scenarios, has been widely studied (Wu et al., 2024; Jin et al., 2024). Our results show that behavior happens in citation generation. We dive into two different types of scenarios where the difficulties of the question or the level of knowledge grasp differ. The distribution of the dataset is shown in Figure 11.

As shown in Figure 6, we separate the dataset into (1) with documents that exactly contain the answer string (Simple); (2) with documents that entail the answer but do not contain the answer string (Hard); (3) $\overline{\textbf{GT}}$. We find the model prefers citing internal knowledge when external documents become harder to leverage.

We separate the dataset by the model's knowledge level to the question into three: Without Knowledge, Low and High-level grasp. The model prefers to cite external knowledge when it has no knowledge about the question and when its knowledge level is high, as demonstrated in Figure 7.

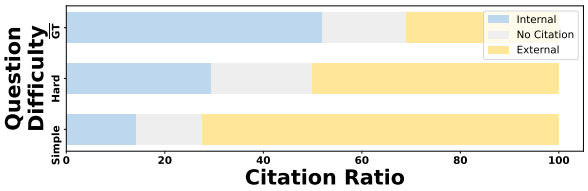

Figure 6: Citations questions with different difficulties in leveraging external knowledge.

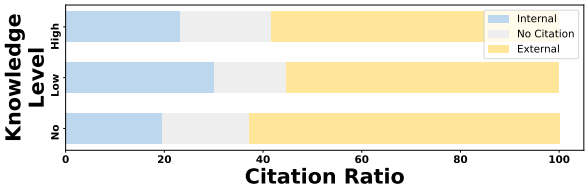

Figure 7: Citations for questions that the Llama3.1-8B has different levels of knowledge.

The tug-of-war between knowledge suggests that the model should rely more on external references when the question is straightforward, and the model's knowledge is limited while leveraging prior knowledge more if the question is complex.

5.4.3 Dishonest internal reference generation

When external documents are not convincing, the model may rewrite them and claim them as internal knowledge. We believe this reflects dishonesty when generating internal references, but we do not penalize this behavior since high-quality rewriting would be more helpful to users. To study this behavior, we define **plagiarism**, which refers to the case as internal reference R_i entails the answer a for questions in the **GT**, $\overline{\textbf{PK}}$ set. The plagiarism rate is $\text{PR} = \frac{1}{N}\sum_{i=0}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\phi(R_{ij}^{in}, a_i)$, and the severity is $\text{PS} = \frac{1}{N}\sum_{i=0}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}P_{ij}\cdot\phi(R_{ij}^{in}, a_i)$.

N is the size of the **GT**, $\overline{\textbf{PK}}$ set and m_i is the number of internal references for the i-th sample. R_{ij}^{in} and P_{ij} are the j-th internal reference of the i-th sample in the **GT**, $\overline{\textbf{PK}}$ set. PR indicates the proportion of plagiarized references, while PS represents the average confidence falsely reported. We show PR and PS over 3 methods in Table 2.

We identify the presence of plagiarism, and our methods demonstrate a relatively lower plagiarism

	Llama3.1-8B		Llama3.1-70B	
	PR ↓	PS ↓	PR ↓	PS↓
PostCite	0.154	0.124	0.128	0.101
Guided-RAE	0.143	0.119	0.111	0.92
Ours	**0.054**	**0.049**	**0.105**	**0.076**

Table 2: PR and PS for different methods.

rate and plagiarism severity, indicating the effectiveness of the alignment process.

Besides, we find that the larger model is worse than the smaller model with RAEL. The phenomenon that a larger model has a higher plagiarism rate might indicate that the model learns more about the preference for convincingness and conciseness and becomes more inclined to rewrite the external documents when the quality of the external documents is not satisfying.

5.5 Ablation of our method

In addition to the Guided-RAEL, we provide results of extra ablations, including (1) Directly generating references and answers without using our RAEL paradigm; (2) Our alignment algorithm without RAEL paradigm; (3) Our alignment algorithm without weighted loss. The results in Tables 3 and 4 are presented, along with an analysis below.

Scenario	Method	Acc	Recall	Conv.	Conc.	ECE
GT, PK	Direct	68.03	30.45	3.44	3.70	0.17
	Ours w/o RAEL	75.87	31.56	3.54	3.76	0.13
	Ours w/o Weight	71.43	50.36	3.61	3.99	0.14
	Ours	**75.90**	**63.72**	**3.61**	**4.05**	**0.10**
GT, $\overline{\text{PK}}$	Direct	53.90	25.11	3.50	3.65	0.18
	Ours w/o RAEL	64.29	22.03	3.52	3.72	0.11
	Ours w/o Weight	69.05	40.17	3.62	**3.85**	0.14
	Ours	**69.05**	**51.84**	**3.64**	3.78	**0.11**
$\overline{\text{GT}}$, PK	Direct	36.40	29.17	3.41	3.20	0.20
	Ours w/o RAEL	61.38	52.79	3.65	3.67	0.13
	Ours w/o Weight	62.24	49.80	**3.66**	3.55	0.12
	Ours	**62.64**	**59.80**	3.65	**3.83**	**0.11**

Table 3: Ablation results on Llama3.1-8B-Instruct

Scenario	Method	Acc	Recall	Conv.	Conc.	ECE
GT, PK	Direct	78.82	40.24	3.48	4.21	0.22
	Ours w/o RAEL	72.01	58.32	3.50	3.92	0.12
	Ours w/o Weight	85.71	58.30	3.65	4.02	0.10
	Ours	**85.72**	**78.79**	**3.69**	**4.42**	**0.10**
GT, $\overline{\text{PK}}$	Direct	61.71	42.76	3.50	4.19	0.22
	Ours w/o RAEL	65.97	53.77	3.54	4.07	0.15
	Ours w/o Weight	**71.88**	53.38	3.44	3.98	**0.12**
	Ours	71.25	**60.32**	**3.60**	**4.47**	0.13
$\overline{\text{GT}}$, PK	Direct	47.80	42.99	3.49	4.10	0.15
	Ours w/o RAEL	62.48	60.12	3.46	4.25	0.12
	Ours w/o Weight	75.11	72.45	**3.67**	4.10	0.12
	Ours	**75.64**	**75.71**	3.67	**4.42**	**0.09**

Table 4: Ablation results on Llama3.1-70B-Instruct

RAEL and weighted loss significantly improve the citation recall, convincingness and conciseness. The accuracy is also slightly improved. The result aligns with our focus on citations in the training and data creation process.

The training without weights sometimes achieves comparable or better results than the weighted training, but that does not happen very often. Considering the low extra cost and convenience of integrating weighted loss, it is still a beneficial way to improve overall performance. We also notice this case happens often in the $\overline{\text{PK}}$ set (with questions for which LLMs have no internal knowledge), which means the weighted loss might be more useful for improving internal citation quality.

6 Human Evaluations

We design a webpage for human evaluations, as shown in Figure 8, and conduct human evaluations to justify our automatic metrics. Three participants with fluent English proficiency took part in the evaluation, each of whom underwent preliminary testing and was assigned 100 examples.

NLI Accuracy. Evaluations demonstrate that our accuracy metric has only a 2.65% False Positive Rate and 6.19% False Negative Rate.

Convincingness aligns with trustworthiness, and Conciseness aligns with verification difficulty. The participants were required to rate the level of trustworthiness and difficulties in verification for certain generated references. We use Pearson Correlation Coefficients (PPCs) as an indicator of the correlation. Table 5 implies that the correlation between human judgment and automatic evaluations is similar to that between individuals.

Evaluator	Convincingness	Conciseness
Automatic	0.53	0.66
Individual β	0.58	0.74

Table 5: PPCs between individual α and two evaluators: (1) Our automatic evaluator and (2) Individual β.

7 Conclusion

In this paper, we present a Context-Prior Augmented Citation Generation task, requiring LLMs to generate citations and fine-grained references from both external contexts and internal parameter knowledge. Our comprehensive evaluation of answer helpfulness, citation faithfulness, and reference trustworthiness reveals the challenge for LLMs to generate trustworthy citations.

We also propose RAEL paradigm for this task and a method, INTRALIGN, to unleash the model's capacity to cite parameter knowledge with trustworthiness, facilitating a more transparent citation generation. In-depth studies reveal the significance of the quality of knowledge sources and highlight the LLM's selective utilization of external and internal knowledge in citation generation.

8 Limitations

Using more open-source models can still enrich our experiments, as different parameter knowledge in the LLMs makes a huge difference in the data sampling and dataset-splitting process.

While we believe our datasets closely reflect the distribution found in real-world scenarios, possible bias may still be introduced during reranking and selection in the data sampling process. We also fabricate Reddit-style data from Wikipedia, which may not represent an authentic low-quality knowledge source. We still observe minor correct answers less than 5% in the $\overline{\textbf{GT}}$, $\overline{\textbf{PK}}$ set, implying that our annotation process still has some omissions.

Although our method is effective, we still leave room to consider and explore the internal mechanism of parameter knowledge utilization, and future works may focus on the intended control of citing external or internal knowledge.

9 Ethical Considerations

We require large models to memorize parameter knowledge, which could raise copyright issues, as some of the data used for training may be copyrighted. However, the complete memorized knowledge does not exceed 200 words per article. Our experimental results do not need full access to the memorized content generated in the model's intermediate steps, and we only analyze the final cited span, which is less than 50 words.

Our dataset includes Reddit-style rewritings of Wikipedia content, which might contain inaccurate or misleading information. We ensure this dataset is private only for our supplementary experiment and not publicly available.

Three non-paid volunteers participated in our human evaluation. We ensured that all participants were informed about the research objectives, the tasks involved, and their role in the evaluation process. Their involvement was entirely voluntary. Volunteers gave their consent and had the right to withdraw at any time without any consequences. We ensured the anonymity and confidentiality of all personal data in the evaluation.

10 Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. U24A20335, No. 62176257). This work is also supported by the Youth Innovation Promotion Association CAS.

References

AI@Meta. 2024. Llama 3 model card.

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.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*, arXiv:2410.08414.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Qiang Ding, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2024. Attention with dependency parsing augmentation for fine-grained attribution. *Preprint*, arXiv:2412.11404.

Yifan Ding, Matthew Facciani, Amrit Poudel, Ellen Joyce, Salvador Aguinaga, Balaji Veeramani, Sanmitra Bhattacharya, and Tim Weninger. 2025. Citations and trust in llm generated responses. *Preprint*, arXiv:2501.01303.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *Preprint*, arXiv:2403.04696.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand. Association for Computational Linguistics.

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.

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*.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024a. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024b. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113, Bangkok, Thailand. Association for Computational Linguistics.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2024c. Enhancing large language models' situated faithfulness to external contexts. *Preprint*, arXiv:2410.14675.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *Preprint*, arXiv:2402.14409.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *Preprint*, arXiv:2409.12941.

Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuiseok Lim. 2023. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:2309.06384*.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024. Improving attributed text generation of large language models via preference learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5079–5101, Bangkok, Thailand. 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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2024. Controllable context sensitivity and the knob behind it. *Preprint*, arXiv:2411.07404.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *Preprint*, arXiv:2410.03727.

Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. 2024. VeraCT scan: Retrieval-augmented fake news detection with justifiable reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 266–277, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2024. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. *Preprint*, arXiv:2412.20309.

Alessandro Scirè, Andrei Stefan Bejgu, Simone Tedeschi, Karim Ghonim, Federico Martelli, and Roberto Navigli. 2024. Truth or mirage? towards end-to-end factuality evaluation with llm-oasis. *Preprint*, arXiv:2411.19655.

Sagi Shaier, Ari Kobren, and Philip V. Ogren. 2024. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17226–17239, Miami,

Florida, USA. 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.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8211–8227, Miami, Florida, USA. Association for Computational Linguistics.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust LLMs: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5984–5996, Bangkok, Thailand. Association for Computational Linguistics.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2024. Search engines in an ai era: The false promise of factual and verifiable source-cited responses. *Preprint*, arXiv:2410.22349.

Fei Wang, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro Szekely. 2021. Retrieving complex tables with multi-granular graph representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1472–1482.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Theodora Worledge, Tatsunori Hashimoto, and Carlos Guestrin. 2024. The extractive-abstractive spectrum: Uncovering verifiability trade-offs in llm generations. *Preprint*, arXiv:2411.17375.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024a. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *The Web Conference 2024*.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2024b. Aliice: Evaluating positional fine-grained citation generation. *Preprint*, arXiv:2406.13375.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. 2024. Crag – comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the parametric knowledge of language models: A unified framework for attribution methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8186, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Yige Yuan, Bingbing Xu, Hexiang Tan, Fei Sun, Teng Xiao, Wei Li, Huawei Shen, and Xueqi Cheng. 2024. Fact-level confidence calibration and self-correction. *Preprint*, arXiv:2411.13343.

Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. 2024a. Verifiable by design: Aligning language models to quote from pre-training data. *CoRR*, abs/2404.03862.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. 2024b. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439, Tokyo, Japan. Association for Computational Linguistics.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *Preprint*, arXiv:2409.10102.

A Related Work

LLM Citation Generation. ALCE (Gao et al., 2023b) introduced a paradigm for citation generation in LLMs and established key evaluation metrics. Subsequent work improved citation quality by refining granularity (Xu et al., 2024b; Zhang et al., 2024b), enhancing model attribution evaluation (Yu et al., 2024), and exploring user-centered effectiveness measure (Worledge et al., 2024). Recent advances follow two approaches: fine-tuning models (Huang et al., 2024b; Ye et al., 2024; Huang et al., 2024a; Li et al., 2024) and designing structured pipeline (Zhang et al., 2024a; Lee et al., 2023; Xu et al., 2024a; Gao et al., 2023a; Ding et al., 2024; Fierro et al., 2024; Sun et al., 2024). Despite these efforts, existing studies have not considered integrating LLMs' internal and external knowledge with confidence quantification.

LLM Parameter Knowledge. Some studies leverage parameter knowledge to enhance generation and attribution (Yu et al., 2023; Sun et al., 2023), while others explore the model's behavior in different scenarios where parameter knowledge and contextual knowledge conflict (Minder et al., 2024; Ming et al., 2024; Cheng et al., 2024). However, these studies have not fully addressed the adaptive use of parameter and contextual knowledge in large models, nor have they provided sufficient interpretability in utilizing parameter knowledge.

Trustworthiness and factuality of LLMs. Some studies have focused on human-centered LLMs, considering the trustworthiness and verifiability of model-generated content (Venkit et al., 2024; Ding et al., 2025). However, these approaches lack quantitative definitions and targeted improvements. To improve the factuality issues in LLM outputs, some studies have developed fact verification methods to detect hallucinations (Min et al., 2023; Niu et al., 2024; Scirè et al., 2024), which have also been used to calibrate model output confidence (Fadeeva et al., 2024; Yuan et al., 2024). Meanwhile, studies have also explored the model's confidence of the generated content in RAG or plain QA tasks (Ozaki et al., 2024; Tao et al., 2024). However, these works have not considered integrating citation tasks to enhance interpretability.

B Human Evaluation

We designed a webpage for conducting human evaluations, as shown in Figure 8. For reference evaluation tasks, each participant is asked to evaluate the Convinvingness and conciseness of a given reference. For answer evaluation tasks, each is asked to evaluate the correctness of the answer.

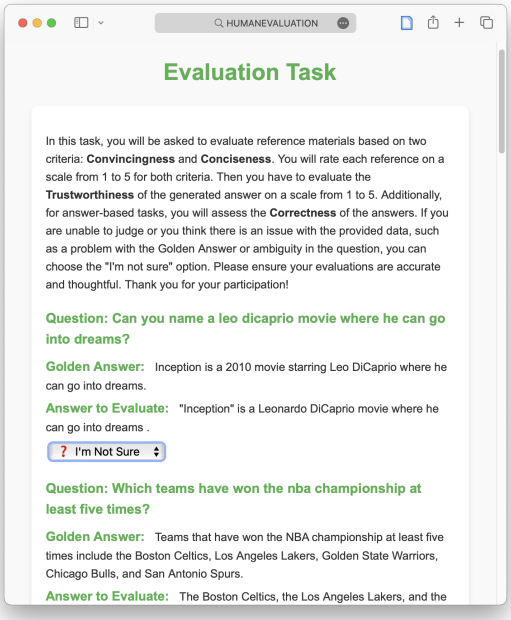

Figure 8: Webpage for human evaluations

NLI Accuracy is better than Exact Match. Previous citation generation tasks use String Exact Match to compute accuracy. However, it is difficult to exhaust all possible answers, and the response may still mention the golden answer even if the intended answer is the other. To alleviate the problems, we desgin our NLI-based Accuracy.

We asked our participants to annotate the answer manually and calculate the False Positive Rate and False Negative Rate of the results from NLI Accuracy and Exact Match. We show the FP Rate, FN Rate and Accuracy in Table 6, which demonstrates our method has relatively fewer mistakes and a higher overall accuracy.

	FP Rate ↓	FN Rate ↓	Accuracy ↑
String Exact Match	3.54%	11.24%	83.11%
Ours	**2.65%**	**6.19%**	**91.27%**

Table 6: Comparison of Exact Match and our method for accuracy evaluation

Convincingness and Conciseness. We visualize the correlation between human evaluations from individual α and our GPT-4o-mini automatic evaluator in Figures 9 and 10, which shows the

correlation on Convincingness and Conciseness, respectively.

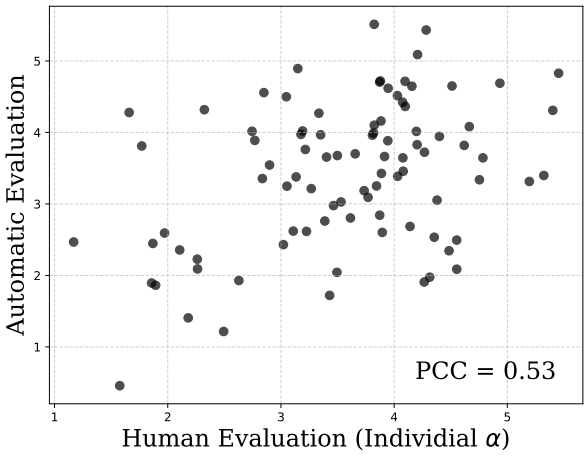

Figure 9: Scatter plot for Convincingness Evaluation. Gaussian noise $\mathcal{N}(\mu = 0, \sigma^2 = 0.5)$ has been added to the points to prevent overlap.

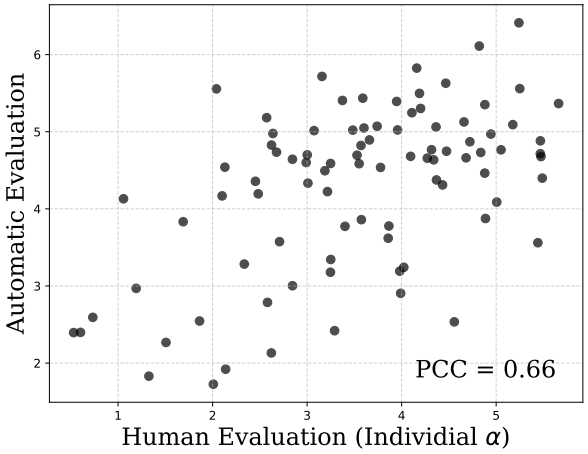

Figure 10: Scatter plot for Conciseness Evaluation. Gaussian noise $\mathcal{N}(\mu = 0, \sigma^2 = 0.5)$ has been added to the points to prevent overlap.

In both plots, the PCC values highlight the correlation between human ratings and the automatic evaluation method.

C Robustness to Shortcut Cases

We discuss three possible shortcut cases: (1) Cite all the documents provided to ensure the Recall score. In this case, references will suffer from high redundancy. (2) Only cite the minimum span to ensure conciseness; the convincingness of the reference will be relatively lower. (3) Only cite external documents or parameter knowledge. In this case, the accuracy metric will significantly decrease on a dataset containing questions the LLM has no knowledge about or documents containing the answer is not provided.

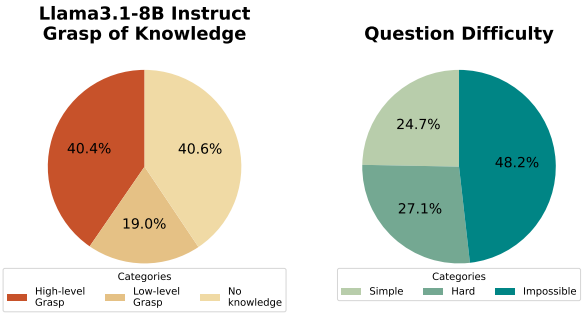

Figure 11: Dataset Distribution

D Datasets

The datasets we used focus on various types of QA, multi-document reasoning, and internal/external knowledge fidelity, making them suitable for comprehensively evaluating the model's overall performance on our task. They are all factual questions from the real world, and relevant knowledge can be retrieved from Wikipedia. The distribution of the dataset according to the partitioning method in 5.4.2 is in Figure 11, and examples are in Table 7.

CRAG (Yang et al., 2024) is a factual question-answering benchmark of 4K question-answer pairs. CRAG is designed to encapsulate diverse questions across 5 domains and 8 question categories, as indicated in Appendix D. To prevent conflicts between external documents and the LLM's internal knowledge due to knowledge updates, we have removed time-sensitive data, retaining only questions categorized as *static*.

FRAMES (Krishna et al., 2024) is a high-quality evaluation dataset designed to test LLMs' ability to provide factual responses and evaluate the reasoning required to generate final answers. Each question in FRAMES requires reasoning with 2-15 Wikipedia articles. We select all the questions that require exactly 2 Wikipedia articles to facilitate the document annotation step.

SituatedFaithfulnessEval (Huang et al., 2024c) benchmarks LLM's ability to demonstrate situated faithfulness, dynamically calibrating their trust in external information based on their confidence in the internal knowledge and the external context. We select the factual QA questions from the **ClashEval** (Wu et al., 2024) subset, which is used to quantify the tug-of-war between an LLM's internal prior and external evidence.

For annotation, we first combine the three datasets, and for each question in the combined dataset, our pipeline first retrieves top-100 passages from a chunked Wikipedia snapshot using a GTR retriever (Wang et al., 2021) and the ques-

Dataset	Category	Example	Amount
CRAG	Simple	Which movie won the Oscar Best Visual Effects in 2021?	580
	Simple with condition	What is a movie to feature a person who can create and control a device that can manipulate the laws of physics?	260
	Multi-hop	How long is the longest river in Alabama?	191
	Comparison	Which movie was created first, A Walk to Remember or The Notebook?	123
	False premise	When did Hamburg become the biggest city in Germany?	100
	Post-processing	How many 3-point attempts did Steve Nash average per game in seasons he made the 50-40-90 club?	48
	Aggregation	How many family movies were there that came out in 1994?	220
FRAMES	Temporal reasoning	Was the person who served as president of the Scottish National Party from 1987 to 2005 alive when the party was founded?	49
	Multiple constraints	As of August 3, 2024, which rabbi worked for both Reform Congregation Keneseth Israel in Philadelphia and Congregation Beth Israel in West Hartford, Connecticut?	124
	Numerical reasoning	What painting was stolen from The Louvre exactly 56 years before the birth of activist and songwriter Serj Tankian?	56
	Tabular reasoning	What is the birthplace and hometown of the winning goal scorer of the 2010 Vancouver Olympics, Men's Ice Hockey event?	42
	Post processing	This athlete was the first man to run the 100 meters in under 10 seconds at an Olympic Games. Which NFL team was he drafted by?	20
SFE	Years	In which year did Godwin Obaseki switch from the APC to the PDP?	480
	Names	Who is the brother of southern gospel music singer Lynda Tait Randle that is associated with the musical groups DC Talk and Newsboys	104
	Locations	Which city is the Saint Nicholas Monastery, an Eastern Orthodox monastery that was made the seat of the Eastern Orthodox Eparchy of Mukachevo in 1491, located in?	142
	News	How much is Sheldon Rankins' contract with the Cincinnati Bengals worth in millions of US dollars as agreed upon on March 30, 2023?	299

Table 7: Examples for each category in the dataset

tion as the query. Then, we apply an NLI model (Honovich et al., 2022) is applied to annotate documents. For each document d and the answer a, if $\phi(d, a) = 1$, then the document is annotated as ground truth. After filtering out the questions without ground truth documents retrieved, we make two data points for each question with different settings, one with a random number of ground truth documents and the other without any ground truth document. Finally, we supplement retrieved documents with $\phi(d, a) \neq 1$ as irrelevant documents so that each data point is paired with exactly 5 documents. According to whether it contains a **G**round **T**ruth document, we split each datapoint into 2 settings **GT** (with ground truth documents) and $\overline{\textbf{GT}}$ (without ground truth documents).

E Implementation Details

NLI model. We apply TRUE as the NLI model, which returns a bool value $\phi(\text{premise}, \text{hypothesis}) = 1$ if the premise entails the hypothesis.

Training setup and Hyper Parameters We use LoRA (Hu et al., 2022) to fine-tune our models with rank $= 8$ and learning rate as lr $= 10^{-4}$. We train 2 epochs on our dataset in total. For GPT-4o and GPT-4o-mini used as baselines and evaluation models, we use temperature $= 0.5$ and top_p $= 0.9$.

E.1 Baselines

We detail the baselines that we used in our experiments below.

Guided-RAEL We use a two-shot prompt to guide the model in using our Rational Attribution and Elaboration framework.

FOOTNOTE. We asked the LLM to generate an answer with reference the `\footnote` format following each sentence. We use `\footnote[confidence]{reference}` to represent internal citation and use `\footnote[idx]{reference}` to represent external citation, where `idx` is the external document index.

POSTCITE. We prompt the LLM to generate

a response according to retrieved documents and segment the response into sentences. For each sentence, we cite the retrieved document with the highest score using a GTR retriever if the score exceeds a given threshold η; otherwise, we ask LLM to generate a document with a confidence score to cite as an internal citation. In our implementation, we dynamically set η to ensure the same total external and internal citations.

RECITATION AUGMENTED GENERATION. Sun et al.'s (2023) work utilizes parameter knowledge to augment LLM Generation. This baseline sample k passages from LLM's parameter knowledge and generates k corresponding answers given each passage as context. The final answer is determined through majority voting. We adopt a logits-based method, CCP (Fadeeva et al., 2024), to obtain a confidence score for each generated passage. In our implementation, we set k to 5. The answer is not a single word, so we are unable to use string match to realize majority voting, so we assume answer a_1 is the same as a_2 if $\phi((q;a_1),(q;a_2))$ or $\phi((q;a_2),(q;a_1))$, and classify all the answers. The final answer is randomly chosen from the largest set.

FRONT. FRONT is a training-based baseline designed to enhance citation quality through fine-grained grounded citations (Huang et al., 2024b). The model first extracts supporting quotes from retrieved documents and uses them to guide the answer-generation process, ensuring precise and grounded responses. Then, it further optimizes the consistency between the grounded quotes and the generated answers using preference optimization techniques. In our implementation, we use our dataset to generate data and train the model following the pipeline of FRONT, replacing our RAEL paradigm and INTRALIGN.

CONTEXTCITE. CONTEXTCITE is a context attribution baseline that identifies the specific parts of the context responsible for generating a model's response (Cohen-Wang et al., 2024). CONTEXTCITE emphasizes contributive attribution, and it achieves this through a surrogate model trained to approximate how excluding or including specific context sources affects the response. CONTEXTCITE utilizes sparse linear modeling to provide efficient and scalable attribution, ensuring that each identified source significantly impacts the generated output. We regard the attributed part in the context as the reference.

F Abstention

When LLM has no knowledge about a question and no ground truth documents are provided, the golden answer is a refusal answer. When the answer contains any of the pre-listed refusal sentences or when the NLI model computes $\phi(a, \text{"Unable to answer."}) = 1$, we consider the model to have given a refusal answer. We measure the rate of refusal in the $\overline{\textbf{GT}}, \overline{\textbf{PK}}$ set of different baselines and our INTRALIGN, as shown in Table 8

Our method allows the model to abstain when necessary. When no ground truth documents are provided and the model also has no knowledge about the question, the model trained on our methods will be more likely to abstain from answering. This behavior allows the model to give a more trustworthy response and reduce hallucinations.

	Llama3.1-Instruct		
	8B	70B	GPT-4o
RECITE†	1.53	1.97	-
FRONT‡	2.55	3.67	-
CONTEXTCITE‡	24.7	22.0	-
FOOTNOTE	18.4	16.19	10.32
POSTCITE	21.4	3.71	17.42
GUIDED-RAE	6.8	6.1	18.7
Ours	**28.84**	**28.21**	-

Table 8: Refusal Rates on test set $\overline{\textbf{GT}}, \overline{\textbf{PK}}$.

G Prompts

We show our prompts for evaluating Convincingness, Conciseness and the prompt for generating RAEL in Multi-scenario Trustworthy Data sampling in Figure 12, 13, and 14, respectively.

```
You are an evaluator tasked with assessing the Convincingness of a text. Convincingness is
↪  defined as the text's ability to avoid raising doubts about its truthfulness in the reader.
↪  Consider the following criteria while scoring:

1. **Logical Consistency**: Evaluate whether the text avoids logical errors or contradictions.
2. **Subjectivity**: Assess whether the language is objective and free from excessive bias or
↪  personal opinions.
3. **Coherence and Focus**: Determine if the arguments are well-connected and focused rather
↪  than scattered or overly parallel.
4. **Information Density**: Consider whether the text provides sufficient relevant information
↪  to substantiate its claims.

Please assign a score between 1 and 5 based on the following detailed guidelines:

1. **Score: 1 (Very Low Convincingness)**
   - Contains multiple logical errors or glaring contradictions.
   - Dominated by subjective or emotional language.
   - Arguments are highly scattered, with no clear connections between points.
   - Lacks sufficient information to support its claims.

2. **Score: 2 (Low Convincingness)**
   - Contains some logical inconsistencies or weak reasoning.
   - Has a noticeable bias or subjective tone.
   - Arguments are somewhat scattered, with limited connections between points.
   - Provides insufficient evidence or relies on vague statements.

3. **Score: 3 (Moderate Convincingness)**
   - Mostly logical with minor inconsistencies.
   - Language is somewhat balanced but may lean towards subjectivity.
   - Arguments are somewhat connected but may lack focus or clarity.
   - Contains adequate but not robust information density.

4. **Score: 4 (High Convincingness)**
   - Logically consistent with no major errors.
   - Language is objective and neutral.
   - Arguments are mostly coherent and focused.
   - Provides substantial and relevant evidence for its claims.

5. **Score: 5 (Very High Convincingness)**
   - Completely free from logical errors or contradictions.
   - Language is fully objective and professional.
   - Arguments are tightly connected and maintain a clear focus.
   - Provides rich, detailed, and highly relevant information to support its claims.

Provide a short explanation of your reasoning, and then output a score between 1 and 5, with
↪  formatting like "Score: 3".
```

Figure 12: Prompt for Convincingness Evaluation

```
You are tasked with assessing the Conciseness of a document sentence by sentence in response to
↪  a given question and answer. For each sentence:
1. Judge whether it positively contributes to answering the question (positive), partially
↪  contributes but feels unnecessary or tangential (neutral), or detracts from the relevance
↪  (negative).
2. Provide a brief explanation for your judgment.

After reviewing all sentences, summarize the overall Conciseness of the document and assign a
↪  score between 1 and 5, following these guidelines:
- **5 (Very High Conciseness):** All sentences are relevant or contribute directly to answering
↪  the question.
- **4 (High Conciseness):** Most sentences are relevant, with a few mildly tangential or
↪  unnecessary.
- **3 (Moderate Conciseness):** A balance of relevant and irrelevant content; reader effort is
↪  moderate.
- **2 (Low Conciseness):** Many sentences are tangential or unnecessary, requiring significant
↪  effort to find relevant information.
- **1 (Very Low Conciseness):** The majority of the document is irrelevant or distracting, with
↪  little useful content.

**Example:**

**Question:** How many championships has Messi won?

**Document:**
1. "Lionel Messi was born on June 24, 1987, in Rosario, Argentina, and is a professional
↪  footballer."
- **Positive:** This sentence establishes Messi as the subject, making it clear the document is
↪  on topic.

2. "His parents are Jorge Messi, a steel factory manager, and Celia Cuccittini, who worked in a
↪  magnet manufacturing workshop."
- **Negative:** This sentence delves into his family background, which feels irrelevant to the
↪  question about championships.

3. "He won his first championship in 2005, leading his team to victory in the U-20 World Cup."
- **Positive:** This sentence is highly relevant, directly addressing Messi's championship
↪  history.

4. "His most recent championship was the 2022 FIFA World Cup, where he captained Argentina to
↪  victory."
- **Positive:** This sentence is also highly relevant, discussing a key championship victory.

5. "Messi hopes to continue playing at a high level and achieve more milestones in his career."
- **Neutral (slightly negative):** While unrelated to his past championships, it serves as a
↪  closing summary and doesn't significantly detract from the document.

**Overall Assessment:**
The document is mostly focused on answering the question, with only one sentence being
↪  significantly off-topic. While the fifth sentence is mildly tangential, it serves as a
↪  conclusion and does not greatly impact the overall relevance.

Score: 4 (High Conciseness)

Provide a short explanation of your reasoning for each sentence, and then output a score
↪  between 1 and 5, with formatting like "Score: 3."
```

Figure 13: Prompt for Conciseness Evaluation

```
You are a Large Language Model with limited knowledge. Given a question, documents, "my knowledge," and a golden answer, please generate a high-quality answer with citation. You should simulate a Large Language Model that thinks step-by-step and outputs references and an answer using the provided documents and "Knowledge in Yourself" (in the "my knowledge section"), but simulate that you cannot see the golden answer. Simulate that you are generating the knowledge yourself, not referring to the "my knowledge" section and the gold answer.

The response needs to follow the following requirements:
1. Your answer should contain all the information in the golden answer provided (i.e., the golden answer is a subset of your full answer).
2. each statement in your answer should be cited properly, with marks like [1] and [2] to indicate the source of the information. When multiple sources are available, cite a minimum set.
3. Your answer should be concise and contain supporting evidence from the documents provided.

Think step by step to generate the full answer by considering the provided `Documents,` `my knowledge>`, and the golden answer. Here is a guidance:
1. Analyze what kind of knowledge you need to answer the question, and try to find supporting evidence in the documents.
2. Use the provided `Documents` first, and if the information is not enough, use "my knowledge" for a supplement. Scrutinize all the possible "my knowledge" and give an appropriate confidence level according to all the possible "my knowledge."
3. Only use `my knowledge` when provided `Documents` are not sufficient. You don't need to use "my knowledge" for comfirming the information in the provided documents or other unnecessary situations.
4. You pretend to be a Large Language Model with limited knowledge, so you can only use the given documents and "my knowledge" to generate the answer. When using "my knowledge", pretend that you are using the knowledge that you have generated yourself. When thinking about my knowledge, use appropriate uncertainty words to indicate the "Confidence provided at the end of "my knowledge" and use 'Internal Knowledge' to mark the source of the knowledge.
5. When citing the provided documents, you should select a fine-grained span from the documents and ensure the span is credible and less redundant. Use Roman numerals to mark the document and use Arabic numerals to mark spans. Use 'Document I' to refer to the first document, and so on.
6. Cite spans using Arabic numerals like [1]. Do not use Roman numerals to cite spans.
7. When using "my knowledge," you should generate a more credible and less redundant version of the knowledge, use Arabic numerals to mark the spans, and output the provided confidence in the last.
8. If none of "my knowledge" is available, admit it honestly and say that it is because of your limited capabilities.
9. If none of the documents and "my knowledge" is relevant to the question, you should still output the steps and an empty reference and then generate an abstention response: "I don't have sufficient knowledge to answer the question, and there is no relevant information in the provided documents to answer the question" with an empty reference.

Here is an example:
<example>


You have to follow the instructions to generate the full answer for the question below:

Question: <question>

`Documents`:
<docs>

`my knowledge`:
<internal_knowledge>

Golden Answer: <golden_answer>

Output:
```

Figure 14: Prompt for Rational Attribution and Elaboration generation