

EARAL: Efficient Attribution of Retrieval-Augmented Answer Generation using Logits

Lucas Albarede¹ Jose G. Moreno¹ Lynda Tamine¹ Luce Lefevre²

¹Université de Toulouse, IRIT, Toulouse, France

²Dir. Technologies Innovation, SNCF, Paris, France

{lucas.albarede,jose.moreno,lynda.tamine}@irit.fr

luce.lefeuvre@sncf.fr

Abstract

Large Language Models (LLMs) generate grounded answers using Retrieval-Augmented Generation (RAG), reducing hallucinations. Prior work has focused on answer and attribution correctness, with recent studies exploring faithfulness through internal *signals* that reflect the model’s decision process. However, these approaches often introduce latency and do not tightly align token generation with attribution. We propose **EARAL**, a novel attribution method for RAG that improves faithfulness through joint answer generation and generation-time attribution with minimal overhead. Documents are marked with special token identifiers, whose logits estimate their contribution during generation and are aggregated into document-level attributions. On the Trust-Align benchmark for trustworthy attributed generation, **EARAL** outperforms state-of-the-art methods across metrics. Additional analysis shows low latency, robustness, and strong alignment with faithfulness.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across NLP tasks, yet they remain prone to hallucination-producing fluent but unsupported or incorrect statements (Ayala and Bechard, 2024; Shuster et al., 2021). Retrieval-Augmented Generation (RAG) mitigates this issue by conditioning generation on retrieved documents (Dao and Le, 2023; Izacard et al., 2023). However, hallucinations persist in RAG systems, particularly in the form of unfaithful or incorrect citations (Krishna et al., 2021; Qi et al., 2024; Wallat et al., 2025; Song et al., 2025). Generating responses that are not only correct but explicitly grounded in supporting evidence is critical in high-stakes domains

Code available at <https://github.com/lucassA/EARAL>

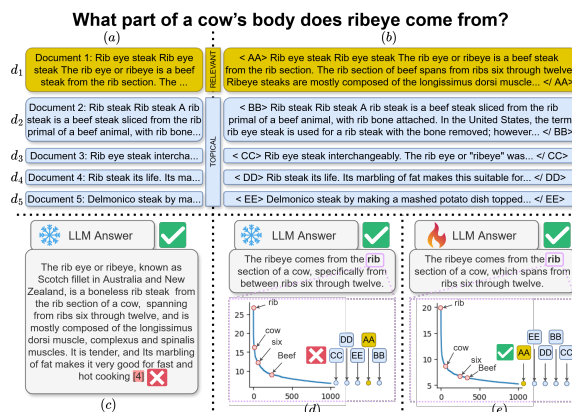


Figure 1: (a) and (c) depict the input and output of a naive RAG model with citations. (d) and (e) show the decreasing logits distribution when generating the token “rib” of a (d) frozen LLM, and a (e) finetuned LLM, **EARAL**, when using our marking presented in (b).

such as healthcare and legal reasoning (Li et al., 2024; Magesh et al., 2024).

To address this challenge, recent work has focused on *document attribution*, which links generated answers to their supporting documents. Existing approaches, however, present important limitations. *Self-generated* citation methods rely on the model to jointly produce answers and citations (Song et al., 2025; Huang et al., 2024a; Xia et al., 2025), often resulting in post-hoc rationalizations rather than faithful reflections of the model’s reasoning. *Retrieval-based* attribution methods use external retrieval or verification steps to assign sources (Gao et al., 2023a; Zhang et al., 2025), but tend to identify plausible documents rather than those directly supporting generation (Wallat et al., 2025). More recently, *model-internal* approaches leverage attention weights, gradients, or ablation signals to better align attribution with the model’s internal dynamics (Qi et al., 2024; Cohen-Wang et al., 2024; Phukan et al., 2024; Yu et al., 2026). Yet these methods either require multiple forward passes or compute additional cross-attention weights, increasing model latency.

In this work, we pursue two key desiderata: (1) *generation-time attribution*, where attribution signals are derived directly from the model’s token-level decision process during generation; and (2) *efficiency*, enabling joint generation and attribution without additional inference passes.

We investigate whether token logits themselves can provide such signals. When retrieved documents are prefixed with unique identifier tokens, the model produces logits for these identifiers at each generation step. We build on recent findings showing that identifier-token logits correlate with document relevance under setwise prompting (Zhuang et al., 2024). This suggests that document-identifier logits may serve as direct estimates of document contribution during answer generation. However, raw logits are affected by selection bias (sensitivity to document order) and token bias (frequency-induced distortions from pre-training) (Wei et al., 2024; Zheng et al., 2024; Jiang et al., 2024), which can undermine faithful attribution.

To address these issues, we propose **EARAL** (Efficient Attribution of Retrieval-Augmented Answer generation using Logits), a *model-internal* attribution framework that jointly fine-tunes answer generation and document attribution. Documents are marked with identifier tokens, and the model is trained to calibrate this model-internal signal, such that their logits reflect grounded document contributions. During inference, attribution scores are obtained directly from the calibrated identifier-token logits at each generation step, requiring no additional forward passes. A subsequent aggregation stage produces statement-level citations. Figure 1 illustrates how debiased identifier logits enable more faithful document ranking compared to a frozen LLM.

We evaluate **EARAL** on ELI5, ASQA, and QAMPARI under the Trust-Align benchmark (Song et al., 2025), which jointly measures answer correctness, grounded citation, and appropriate refusal. Our results show consistent improvements over both state-of-the-art attribution methods and trustworthiness-oriented baselines, while incurring only marginal latency overhead.

Our main contributions are: (1) We introduce a *model-internal* attribution framework that derives document contributions directly from token logits during generation, enabling efficient attribution during generation in RAG; (2) We propose a joint fine-tuning strategy that calibrates identifier-token logits, mitigating selection and token biases; (3) We

provide comprehensive experiments demonstrating improved faithfulness and efficiency on standard long-form QA benchmarks.

2 Related Work

2.1 Attributed Answer Generation

Attribution methods can be broadly grouped into three categories (Ding et al., 2025).

Self-generated approaches treat attribution as part of the generation process, jointly producing answers and citations (Song et al., 2025; Huang et al., 2024a; Xia et al., 2025; Menick et al., 2022). While simple to deploy, they rely on the model to justify its own outputs, often resulting in post-hoc rationalizations with limited faithfulness.

Retrieval-based methods assign citations through additional retrieval or verification steps (Huang et al., 2024b; Gao et al., 2023a; Zhang et al., 2025; Slobodkin et al., 2024; Sancheti et al., 2024). For example, START (Huang et al., 2024b) decomposes answers into claims and retrieves supporting evidence for each. Although more structured and interpretable, these pipelines introduce additional latency and may still select plausible rather than faithfully grounded sources.

Model-internal approaches leverage internal signals such as attention weights, gradients, or distribution shifts to align generated tokens with supporting context (Qi et al., 2024; Cohen-Wang et al., 2024; Phukan et al., 2024; Ding et al., 2025; Yu et al., 2026). MIRAGE (Qi et al., 2024), for instance, identifies sensitive tokens via ablation and applies contrastive feature attribution to estimate document influence. AttnUnion (Ding et al., 2025) aggregates attention weights between prompt and response tokens. By relying on model dynamics, these methods aim for improved faithfulness, but typically require additional computation (e.g., ablation passes) or attribute post-hoc after full answer generation.

In contrast, **EARAL** uses identifier-token logits as a model-internal signal that is explicitly calibrated through supervised fine-tuning. During inference, document contributions are read directly from these calibrated logits during autoregressive generation, enabling generation-time attribution with minimal additional latency.

2.2 Evaluation of Attributed Generation

Attributed generation is commonly evaluated along two axes: answer correctness and attribution qual-

ity. Correctness is measured using semantic-overlap metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang* et al., 2020). Attribution quality has motivated dedicated metrics assessing whether cited documents support generated statements (Gao et al., 2023b; Yue et al., 2023; Djeddal et al., 2024).

ALCE (Gao et al., 2023b), for example, computes citation precision and recall using NLI-based entailment between statements and passages. Yue et al. (Yue et al., 2023) propose ATTScore to evaluate coverage and contradiction. More recently, Trust-Align (Song et al., 2025) highlights the limitations of traditional metrics for RAG and introduces trustworthiness-oriented measures combining correctness, grounded citation, and refusal behavior. We adopt the Trust-Align framework to assess both attribution faithfulness and overall RAG reliability.

2.3 Selection and Token Bias in LLMs

LLMs compute token probabilities based on both input context and internalized training statistics. Recent studies identify *selection bias*-sensitivity to option ordering or positional effects and *token bias*, where certain identifiers receive disproportionate probability mass due to pre-training frequency (Wei et al., 2024; Zheng et al., 2024; Jiang et al., 2024).

These biases are particularly pronounced in selection settings, where the model must choose among labeled options, often favoring specific identifiers independently of content. Since our method relies on identifier-token logits to estimate document contribution, explicitly mitigating these biases is crucial for faithful attribution.

3 Problem Statement

3.1 Preliminaries and notations

Attribution. Let us consider an LLM $\mathcal{M}$ and S a statement generated by $\mathcal{M}$. An attribution a of statement S is composed of a set of document identifiers $\{id_i\}_{i=1}^{n_S}$ citing context document passages $\{d_i\}_{i=1}^{n_S}$ to support or ground statement S . An attribution a can be either model-generated or not. An attributed answer statement S^a is composed of the concatenation of a model-generated statement S and its attribution a .

Generation-time attribution Let us consider an LLM $\mathcal{M}$, C a context composed of documents $C = \{d_i\}_{i=1}^K$, S a statement generated by $\mathcal{M}$ and $a = \{id_i\}_{i=1}^{n_S}$ an attribution of S . We refer to S^a as

using generation-time attribution when attribution a is estimated from *signals* gathered at each step of statement generation, conditioned only on the information available to the autoregressive model at that step (i.e., the query, context, and previously generated tokens). This terminology emphasizes when the signals are collected and does not imply causal identification in the interventionist or structural causal-model sense. In **EARAL**, these generation-time signals are calibrated during training using supervised attribution targets.

3.2 Task definition

The goal of retrieval-augmented attributed answer generation is to answer query q with answer A using a retrieved context C from a corpus of documents D , $C = \{d_1 \dots d_K | d_i \in D\}$. Answer A is composed of a set of attributed answer statements $\{S_1^a \dots S_m^a\}$ where S_i^a is the concatenation of the answer statement S_i and each element of its attribution $a_i = \{id_k \in T\}_{k=1}^{k_i}$. Consider an LLM $\mathcal{M}$ pre-trained over a tokenizer dictionary T , instructed with a query q , and context C . The main problem we tackle in our work consists of targeting faithfulness by producing attributions from generation-time *signals* collected during answer generation.

Formally, we aim at inducing attributions $\{a_i\}_{i=1}^m$ to answer statements $\{S_i\}_{i=1}^m$ generated by $\mathcal{M}$ such that $\{S_i^a\}_{i=1}^m$ are faithfully grounded in the retrieved context.

4 Answer Generation and Attribution

4.1 Method Overview

An overview of **EARAL** is shown in Figure 2. We assume a generative LLM $\mathcal{M}$ with vocabulary V . Given a query q and retrieved context $C = \{d_1, \dots, d_K\}$, the model generates an answer $A = \{S_1, \dots, S_m\}$ composed of statements $S_i = (s_{i1} \circ \dots \circ s_{in_i})$, where $s_{ij} \in V$.

Generation is autoregressive: at each step, the next token s_{ij} is sampled from

$$P(V | q, C, A_{S_{<i}}, S_{i<j}, \mathcal{M})$$

where $A_{S_{<i}}$ denotes previously generated statements and $S_{i<j}$ the preceding tokens in S_i . We denote by $\mathcal{M}(s_{ij} | q, C, A_{S_{<i}}, S_{i<j})$ the pre-softmax *token logit* associated with s_{ij} .

Building on prior findings that identifier-token logits correlate with document relevance (Zhuang et al., 2024), we use document identifiers to obtain

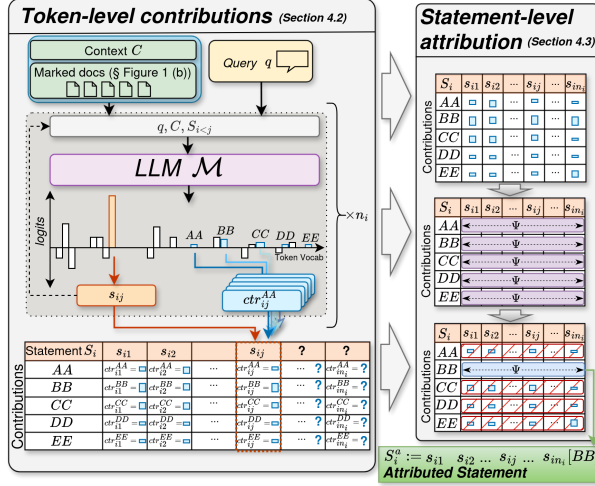


Figure 2: Overview of our proposal, **EARAL**. Given a query and a context composed of documents, the first step (left) consists in computing token-level contributions of marked documents with specific tokens and collecting the logits associated with these tokens during generation. Once a statement has been generated, a second step (right) aggregates the computed contributions into statement-level attribution.

generation-time attribution signals during generation. Each document $d_k \in C$ is prefixed with a unique identifier token $id_k \in V$, producing a marked context $C = \{md_1, \dots, md_K\}$.

At each generation step of token s_{ij} , the logit of identifier id_k , denoted l_k , serves as a token-level contribution score of document d_k :

$$ctr_{ij}^k = \mathcal{M}(id_k | q, C, A_{S_{<i}}, S_{i<j})$$

To mitigate token and selection biases (Wei et al., 2024; Zheng et al., 2024), we fine-tune $\mathcal{M}$ jointly for answer generation and logit calibration, yielding a debiased model $\mathcal{M}_{aa}$ such that

$$ctr_{ij}^k = \mathcal{M}_{aa}(id_k | q, C, A_{S_{<i}}, S_{i<j})$$

Finally, token-level contributions are aggregated into statement-level attributions using an aggregation function Ψ :

$$a_i = \Psi\left(\{ctr_{ij}^k\}_{j=1}^{n_i}\right)_K^K$$

4.2 Learning token-level contribution

4.2.1 Document marking

Marking functions are one of the core components of **EARAL**. A *marking* function prefixes the content of documents in context C with identifier tokens. We investigate three *marking* functions, presented in Table 1. Each function transforms each

document d in the context C into a marked document md_i agnostically to the nature of the identifier token id_i :

- $md_i = marking_{BA}(d_i, id_i)$: we incorporate the identifier token id_i **B**efore and **A**fter the document d_i using greater-than ($>$) and less-than ($<$) signs. The objective of this strategy is to leverage the LLM’s capabilities of processing code-like structure, motivated by the huge amount of code data present in LLM’s training phases.
- $md_i = marking_{BAS}(d_i, id_i)$: we incorporate the identifier token id_i **B**efore and **A**fter each **S**entence of the document d_i using greater-than and less-than signs. This is a similar yet finer-grained strategy than the one above.
- $md_i = marking_{AW}(d_i, id_i)$: we incorporate the identifier token id_i before **A**ll **W**ords of document d_i . The objective of this strategy is to study whether repetition allows the LLM to yield higher logits for the identifier tokens during generation.

4.2.2 Debiasing LLM’ token logits to learn document contribution

Motivation. Our objective is to induce document attribution during answer generation by leveraging the logits of document-identifier tokens. However, prior work has shown that LLMs exhibit biases that can undermine this strategy (Zheng et al., 2024; Wei et al., 2024).

First, *selection bias* makes models sensitive to the position of information in the context. In our setting, identifier-token logits could therefore reflect document rank rather than actual content relevance. Second, *token bias* causes models to assign disproportionate probability mass to certain tokens due to their pre-training frequency. This may artificially inflate or suppress identifier-token logits, leading to unreliable attribution, particularly when relying on small probability differences.

A straightforward mitigation is prompt ablation, i.e., removing documents and measuring the change in generation probabilities (Qi et al., 2024; Cohen-Wang et al., 2024). However, this requires multiple forward passes and substantially increases latency (§6.3.1).

Finally, our problem is structurally constrained: we seek to calibrate only a small subset of logits corresponding to document identifiers ($K \ll |T|$),

Marking function	Marked documents
$marking_{BA}$	$\langle AA \rangle$ The rib eye or ribeye is a beef steak from the rib section. The rib section of beef spans from ribs six through twelve. $\langle AA \rangle$ $\langle BB \rangle$ A rib steak is a beef steak sliced from the rib primal of a beef animal. $\langle BB \rangle$
$marking_{BAS}$	$\langle AA \rangle$ The rib eye or ribeye is a beef steak from the rib section. $\langle AA \rangle$ $\langle AA \rangle$ The rib section of beef spans from ribs six through twelve. $\langle AA \rangle$ $\langle BB \rangle$ A rib steak is a beef steak sliced from the rib primal of a beef animal. $\langle BB \rangle$
$marking_{AW}$	AA The AA rib AA eye AA or AA ribeye AA is AA a beef AA steak AA from [...] AA spans AA from ribs AA six AA through AA twelve. BB A BB rib BB steak BB is BB a BB beef BB steak BB sliced BB from BB the BB rib BB primal BB of BB a BB beef BB animal.

Table 1: Illustration of the marking strategies where two documents are marked with identifier “AA” and “BB”.

while preserving the model’s predictive distribution over answer tokens. To address these challenges, we fine-tune a backbone LLM $\mathcal{M}$ jointly for answer generation and attribution.

Document contribution through logits learning.

To mitigate the aforementioned biases, we fine-tune the backbone LLM $\mathcal{M}$ jointly for answer generation and attribution using a training set $\mathcal{D}_{train}$ of triplets $(q, C, \hat{A})$, where $\hat{A} = \{\hat{S}_1^a, \dots, \hat{S}_m^a\}$ is a gold attributed answer and each $\hat{S}_i^a$ concatenates an answer span $\hat{S}_i$ with its document identifiers $\hat{a}_i = \{id_k \mid d_k \in C\}$.

We optimize the attributed answer loss:

$$\mathcal{L}_{aa} = \sum_{(q, C, \hat{A}) \in \mathcal{D}_{train}} \sum_{\hat{S}_i \in \hat{A}} (1 - \alpha) \mathcal{L}_{ans} + \alpha \mathcal{L}_{att}, \quad (1)$$

where α balances generation and attribution.

The answer loss is the standard next-token cross-entropy:

$$\mathcal{L}_{ans} = \sum_{\hat{s}_{ij} \in tok(\hat{S}_i)} -\log p_{ij}, \quad (2)$$

with $p_{ij} = p(\hat{s}_{ij} \mid q, \hat{A}_{S_{<i}}, S_{i<j}, C, \mathcal{M})$

The attribution loss is an MSE objective (Jiao et al., 2020; Kim et al., 2021) applied only to document-identifier logits:

$$\mathcal{L}_{att} = \sum_{\hat{s}_{ij} \in tok(\hat{S}_i)} \sum_{k=1}^K (l_k - \hat{l}_k)^2 \quad (3)$$

where $l_k = \mathcal{M}(id_k \mid q, \hat{A}_{S_{<i}}, S_{i<j}, C)$ and $\hat{l}_k$ are scaled logit targets encoding levels of groundedness (§ Appendix C).

At inference, the contribution of document d_k to token s_{ij} is directly obtained from the learned identifier logit:

$$ctr_{ij}^k = \mathcal{M}_{aa}(id_k \mid q, A_{S_{<i}}, S_{i<j}, C). \quad (4)$$

This yields attribution signals aligned with the model’s token-level decision process, without additional forward passes.

4.3 From token- to statement-level attribution

The objective here is to induce for each model-generated statement $S_i(s_{i1} \circ \dots \circ s_{in_i})$ an attribution $a_i = \{id_k\}_{k=1}^{n_i}$ composed of document-identifier tokens. To this end, we mainly rely on the token-level contributions ctr_{ij}^k computed at each generated token s_{ij} , which we aggregate using an aggregation function Ψ , such as $a_i = \Psi \left[\{ctr_{ij}^k\}_{j=1}^{n_i} \right]_{k=1}^K$. The contribution scores ctr_{ij}^k are inherently non-discrete, allowing for a finer understanding of the attribution mechanism and improving interpretability. However, user-oriented systems benefit from having discrete attributions as it is less tedious to process for a human. To be more aligned with this requirement, we use a contribution aggregation function Ψ , which transforms the token-level contributions into a boolean attribution, applicable to each document d_k in the context C and returning 1 if d_k is to be included in a_i , 0, otherwise. Specifically, we define Ψ as follows:

$$a_i = \{id_k \mid \sum_{j=1}^{n_i} \mathbf{1}_{ctr_{ij}^k > \phi_{prop}} > \lambda(n_i)\}_{k=1}^K \quad (5)$$

where $\mathbf{1}_{ctr_{ij}^k > \phi_{prop}} = 1$ if $ctr_{ij}^k > \phi_{prop}$ and 0 otherwise; ϕ_{prop} is a contribution threshold and λ is a token proportion threshold. Similar to top-k pooling operator (Kalchbrenner et al., 2014), the underlying intuition is that the documents attributed to the answer are those that contribute highly to at least a thresholded percentage of the associated tokens.

5 Experimental Setup

Benchmark and Metrics. We evaluate under the *Trust-Align* framework (Song et al., 2025), which extends ALCE (Gao et al., 2023b) to jointly assess answer correctness, citation groundedness, and appropriate refusal. Experiments are conducted on QAMPARI (Amouyal et al., 2023), ASQA (Stelmakh et al., 2022), and ELI5 (Fan et al., 2019).

We report F1AC (answer correctness), F1GR (refusal), F1GC (grounded citation), and TRUST (their mean), following (Song et al., 2025). We

additionally report average latency per query. Full metric definitions are provided in Appendix A.1.

Baselines. We compare **EARAL** to:

(i) *Trust-Align* baselines from (Song et al., 2025), including PostCite*, PostAttr*, FRONT*, and Trust-Align fine-tuned; (ii) *ContextCite* (Cohen-Wang et al., 2024) which is a recent *model-internals* approach; (iii) *MIRAGE* (Qi et al., 2024), a faithful attribution method based on input ablations; (iv) *Llama 3.1 8B in zero-shot* and *Trust-Align fine-tuned* variants.

Unless otherwise stated, all methods use Llama 3.1 8B and the ALCE top-5 retrieved documents. Baselines details can be found in Appendix B.

Training Data. Training combines Trust-Align, HAGRID (Kamalloo et al., 2023), and ExpertsQA (Malaviya et al., 2024). We retain examples with attributable spans and augment each dataset with 25% refusal cases to enforce robustness to missing evidence. The final training set contains 20,312 examples. Detailed preprocessing steps are provided in Appendix A.2.

Implementation Details. **EARAL** relies on identifier-token marking and logit supervision for document attribution. Attribution is aggregated at the sentence level, with a refusal fallback when no document is attributed for any answered sentence.

Models are fine-tuned for two epochs with standard optimization settings. Hyperparameters, identifier-token selection, logit target labels, aggregation thresholds, and hardware details are reported in Appendix C.

6 Results and analysis

6.1 Main results

Table 2 presents results on ASQA, QAMPARI, and ELI5, split into baselines (top) and our marking strategies and models (bottom). We can see that **EARAL** competes with state-of-the-art methods, with *marking_{BA}* consistently and significantly improving upon Llama 3.1 8B fine-tuned with *Trust-Align* ($\triangleright$ *Trust-Align*) and *MIRAGE* in TRUST score across all datasets. For example, we gain 4.76 TRUST score over $\triangleright$ *Trust-Align* on ELI5, a 10.7% improvement (48.94 vs 44.18).

We report significance against both baselines because they serve complementary roles: *MIRAGE* is the closest model-internal attribution baseline, while fine-tuned *Trust-Align* provides a strong fine-tuned comparison.

EARAL also significantly improves upon *MIRAGE* in F1GC by up to 56.25% (82.69 vs 52.92), 36.6% (41.91 vs 30.68), and 34.98% (53.28 vs 39.47) on ASQA, QAMPARI, and ELI5, respectively. Additionally, **EARAL** attains higher TRUST scores than *ContextCite* on each dataset.

EARAL outperforms Llama3 8B fine-tuned with *Trust-Align* on all metrics for ASQA and ELI5, improving TRUST by 6.63% (70.86 vs 66.45) and 9.96% (48.94 vs 44.83), respectively. On QAMPARI, we hypothesize the divergence stems from backbone model differences, as $\triangleright$ *Trust-Align* also shows weaker results on this dataset.

For zero-shot prompting, *marking_{BAS}* outperforms *marking_{BA}* and *marking_{AW}* on nearly all metrics, with clear F1GC gains while maintaining similar performance on other metrics (e.g., 20% improvement on QAMPARI: 16.83 vs 13.96). However, with fine-tuning, *marking_{BA}* outperforms all strategies including zero-shot scenarios, suggesting that overly refined marking hurts fine-tuning performance and that simpler marking is more suitable.

Regarding latency, our proposals match citation-generating methods (e.g., $\triangleright$ *Trust-Align*) but are faster than other *model-internals* approaches. **EARAL** with *marking_{BA}* is $17.97\times$ faster than *ContextCite* (335.55 ms vs 6031.12 ms) and $104.38\times$ faster than *MIRAGE* (335.55 ms vs 35024.44 ms). This advantage stems from relying only on existing logit information rather than extra parameters.

6.2 Ablation Study

Table 3 presents our ablation study evaluating the debiasing logit strategy (§ 4.2.2) and the proposed aggregation function Ψ (§ Eq.5). The *w/o debiasing* scenario uses initial model-output logits instead of debiased ones. The *w/o aggregation* scenario employs naive aggregation operators since raw token-based contribution values are required. We consider two alternatives: (1) the *max* strategy $a_i = \{id_k | \arg \max_k (ctr_{ij}^k) > \phi_{max}\}_{k=1}^K$ where ϕ_{max} is a contribution threshold, attributing documents that contribute highly to at least one token (Qi et al., 2024), similar to max-pooling (Shen et al., 2018); (2) the *avg* strategy $a_i = \{id_k | \arg [avg[(ctr_{ij}^k)]_{j=1}^{n_i}] > \phi_{avg}\}_{k=1}^K$ where ϕ_{avg} is a contribution threshold, attributing documents that contribute highly to all tokens, similar to mean-pooling (Reimers and Gurevych, 2019).

Table 3 shows that removing either logit-

		ASQA				QAMPARI				ELIS				Average Latency	
		F1AC	F1GR	F1GC	TRUST	F1AC	F1GR	F1GC	TRUST	F1AC	F1GR	F1GC	TRUST		
Baselines															
Llama3	PostCite* (Gao et al., 2023b)	32.98	53.31	28.01	38.10	6.10	34.52	8.42	16.35	20.80	45.88	8.06	24.91	-	
	PostAttr* (Gao et al., 2023b)	32.98	53.31	5.95	30.75	6.10	34.52	1.64	14.09	20.80	45.88	1.25	22.64	-	
	Fine-tuned w/ FRONT* (Huang et al., 2024a)	62.25	41.62	66.14	56.67	13.53	22.78	20.42	18.91	18.99	17.85	44.69	27.18	-	
	Fine-tuned w/ Trust-Align* (Song et al., 2025)	52.35	66.06	80.95	66.45	33.85	71.11	48.01	50.99	22.57	65.06	46.85	44.83	-	
Llama3.1	ContextCite (Cohen-Wang et al., 2024)	59.81	64.31	60.03	61.38	1.91	63.02	29.19	31.37	29.54	55.88	47.55	44.32	6031.12	
	MIRAGE (Qi et al., 2024)	59.81	64.31	52.92	59.01	1.91	63.02	23.70	29.54	29.54	55.88	39.47	41.63	35024.44	
	Zero-Shot Prompting	59.81	64.31	60.85	61.66	1.91	63.02	21.50	28.81	29.54	52.88	43.77	42.06	339.70	
	▷ Fine-tuned w/ Trust-Align (Song et al., 2025)	53.64	65.33	84.52	67.83	10.69	68.83	51.05	43.53	25.52	67.09	39.91	44.18	329.73	
Our markings and models															
Llama3.1	Zero-Shot Prompting	marking _{BA}	61.61	65.60	44.23	57.15	2.24	63.00	13.96	26.40	30.36	51.71	35.01	39.03	341.22
		marking _{BAS}	62.29	65.85	51.02	59.72	2.51	62.38	16.83	27.24	29.51	52.71	37.17	39.80	347.87
		marking _{AW}	60.36	62.40	36.65	53.14	2.06	62.22	12.54	25.60	29.65	53.09	27.00	36.58	411.89
	EARAL	marking _{BA}	61.52 [▲]	68.35 [▲]	82.69 [■]	70.86[■]	31.73 [■]	69.24 [■]	41.91 [■]	47.63 [■]	27.98 [▲]	65.58 [■]	53.28 [■]	48.94[■]	335.55
		marking _{BAS}	61.08 [▲]	63.76 [■]	71.56 [■]	65.47 [■]	28.01 [■]	68.24 [■]	30.68 [□]	42.27 [■]	27.50	64.45 [■]	41.96 [▲]	44.14 [□]	341.54
		marking _{AW}	54.76	63.44	81.11 [■]	66.44 [■]	26.18 [■]	68.02 [□]	34.55 [■]	42.91 [■]	26.74	60.15 [□]	49.23 [■]	46.23 [■]	407.11

Table 2: Comparison of **EARAL** performance with state-of-the-art at 8B parameters size. Models with a star (*) have their performance measures reported from (Song et al., 2025). ▲ and △ symbols indicate significant improvements over the baseline model ▷ Fine-tuned w/ Trust-Align using a paired t-test with $\rho = 0.01$ and $\rho = 0.05$, respectively. ■ and □ symbols indicate significant improvements over the baseline model ▢ MIRAGE using a paired t-test with $\rho = 0.01$ and $\rho = 0.05$, respectively. Bold indicates best TRUST scores per dataset. Average Latency is reported in ms per query.

debiasing or k-pooling aggregation decreases performance, confirming their joint contribution to **EARAL**. Debiasing is critical across all marking strategies and datasets, with performance drops up to -47.21% when removed. Similarly, replacing our attribution strategy with the *max* operator yields negative effects ranging from 0.69% to 19.16% . The *avg* aggregation behaves dataset-dependently: it slightly benefits *marking_{BAS}* and *marking_{AW}* on ASQA and ELI5, but not QAMPARI, and does not improve our strongest strategy, *marking_{BA}*.

		ASQA	QAMPARI	ELI5
<i>marking_{BA}</i>		70.86	47.60	48.94
	w/o debiasing	55.13 ^{↓22.20%}	25.13 ^{↓47.21%}	30.01 ^{↓38.68%}
	+ max	67.76 ^{↓4.37%}	47.27 ^{↓0.69%}	46.65 ^{↓4.68%}
	+ avg	69.31 ^{↓2.19%}	45.55 ^{↓4.31%}	46.98 ^{↓4.00%}
<i>marking_{BAS}</i>		65.73	39.39	42.54
	w/o debiasing	54.64 ^{↓16.87%}	25.89 ^{↓34.27%}	31.68 ^{↓25.53%}
	+ max	60.17 ^{↓8.46%}	37.53 ^{↓4.72%}	37.97 ^{↓10.74%}
	+ avg	65.80 ^{↑0.11%}	36.26 ^{↓7.95%}	42.23 ^{↓0.73%}
<i>marking_{AW}</i>		68.05	43.34	47.13
	w/o debiasing	56.74 ^{↓16.62%}	26.17 ^{↓39.62%}	33.05 ^{↓29.87%}
	+ max	55.01 ^{↓19.16%}	42.85 ^{↓1.13%}	43.46 ^{↓7.79%}
	+ avg	66.07 ^{↓2.91%}	40.94 ^{↓5.54%}	47.98 ^{↑1.80%}

Table 3: Performance results (TRUST score) of the ablation study. ↓ and ↑ represent respectively the percentage decrease and percentage increase in TRUST score over **EARAL** (w.r.t. each marking strategy).

6.3 EARAL Analysis

6.3.1 Logit debiasing

To evaluate our logit-based debiasing approach, we compare against two alternative scenarios from prior work (Qi et al., 2024; Cohen-Wang et al., 2024): (1) *ablate-repeat*, which ablates parts of the prompt and analyzes effects on generation proba-

	ASQA		QAMPARI		ELI5	
	T	Lat	T	Lat	T	Lat
<i>ablate-repeat</i>	50.36	771.1*	26.53	441.4*	29.99	734.2*
<i>KL-proba</i>	62.01	402.6	39.44	224.4	43.18	381.7
EARAL	70.86	403.3	47.63	227.4	48.94	375.9

Table 4: Analysis of different logit debiasing strategies using *marking_{BA}*. Latency (Lat) is reported in ms per query. * Note that parallel computing of the contributions may reduce the overhead without cancelling it completely. Bold scores indicate the highest value for the TRUST (T) score and lowest value for the latency.

bilities, with contributions redefined as:

$$ctr_{ij}^k = \mathcal{M}(id_k|q, A_{S_{<i}}, S_{i<j}) - \mathcal{M}(id_k|q, A_{S_{<i}}, S_{i<j}, C) \quad (6)$$

and (2) *KL-proba*, which debias token generation probabilities using KL-divergence (Hinton et al., 2015):

$$\mathcal{L}_{att}^{KL} = \sum_{\hat{s}_{ij} \in tok(\hat{S}_i)} \text{KL}(P(\hat{a}_i) \parallel P(id | q, A_{S_{<i}}, S_{i<j}, C, \mathcal{M})) \quad (7)$$

Table 4 presents performance and latency results. Our logit-based strategy in **EARAL** substantially outperforms *ablate-repeat* with TRUST score improvements of 28.93%, 44.29%, and 38.72% on ASQA, QAMPARI, and ELI5 respectively, while requiring nearly half the latency. Compared to *KL-proba*, our approach achieves improvements of 14.27% (70.86 vs. 62.01), 20.76% (47.63 vs. 39.44), and 13.33% (48.94 against 43.18) on the same datasets with comparable latency, demonstrating the effectiveness of operating in logit space.

6.3.2 Robustness analysis of EARAL

We analyze **EARAL**’s robustness to variations in context configuration (detailed explanations in Appendix D.1).

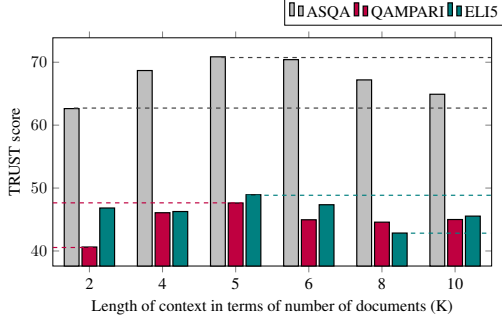


Figure 3: Robustness of **EARL** to the length of context.

Context length. Figure 3 shows performance across varying numbers of documents. While optimal performance occurs at $K = 5$ (matching training), **EARL** remains robust to small variations with TRUST score drops of only 0.63%-5.29% when using 4 or 6 documents.

Document order. **EARL** demonstrates robustness to both identifier token selection and document ordering (Figure 4 for ASQA, other datasets behave similarly §Appendix D.3). Performance converges across random vs. alphabetical token assignment and vanilla vs. reversed document ordering as training progresses, indicating the **EARL** learns position-invariant attribution.

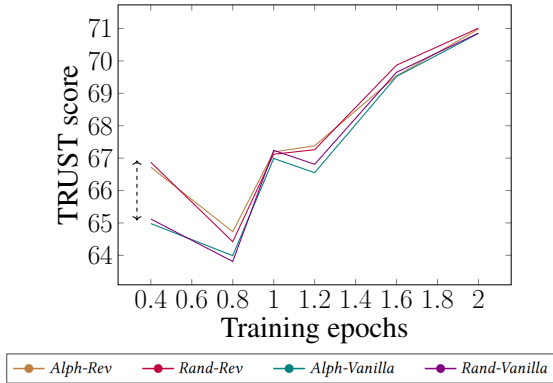


Figure 4: Robustness of **EARL** to document ordering and identifier token selection on ASQA. *Rand-Vanilla* (purple) corresponds to the default configuration of our $marking_{BA}$ strategy.

Dataset	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
ASQA	69.19	69.85	70.86
QAMPARI	46.01	46.78	47.63
ELI5	48.14	49.16	48.94

Table 5: Impact of the proportion threshold (λ) on the Trust score when using the $marking_{BA}$ strategy.

Proportion threshold. Table 5 demonstrates that **EARL** shows minimal sensitivity to the aggrega-

tion threshold λ (§ Eq.5), with TRUST score variations under 3.40% across $\lambda \in \{0.25, 0.5, 0.75\}$ (§ Appendix D.2), suggesting confident and consistent identifier logit predictions.

6.4 Analysis of Faithfulness

Evaluating attribution faithfulness is challenging, but recent work proposes a practical approximation (Wallat et al., 2025). Given a query q , context C , and a pair (S, a) consisting of statement S and attribution $a = \{id_k\}$, an attribution is considered faithful if: (1) altering the source document d_k changes the generated statement or the attribution decision; and (2) adding irrelevant documents to C does not affect attribution as long as the statement remains unchanged.

We conduct a quantitative evaluation of faithfulness attribution on ASQA. We consider three scenarios: (a) a *Baseline* RAG setting with gold attribution; (b) an *Ablation* scenario where the answer-bearing span is removed from the gold document; and (c) a *Replacement* scenario where a context document is substituted with an unrelated one. For each query, we measure the proportion of answer or attribution changes relative to the *Baseline* under both *Ablation* and *Replacement* scenarios. We separate queries where the baseline answer is correct from those where it is incorrect. For incorrect cases, the ablation removes the span supporting **EARL**’s own answer (identified using Llama3.1 8B).

A faithful attribution model should exhibit frequent changes under *Ablation* and minimal changes under *Replacement*. As shown in Table 6, **EARL** follows this pattern: answer/attribution changes occur in 94.80% (correct baseline answers) and 89.16% (incorrect baseline answers) of *Ablation* cases, but only in 1.13% and 3.61% of *Replacement* cases, respectively.

Although limited in scale, these results provide empirical evidence that **EARL** exhibits behavior consistent with faithful attribution. In particular, this intervention-style ablation/replacement analysis provides our strongest evidence that the attributed documents are functionally relevant to the generated answer, beyond the fact that the attribution signals are collected online during generation. A complementary qualitative analysis of **EARL** behavior is presented in Appendix D.4.

		# Qu.	# Ch.	% Changes
Ablation	CA	616	584	94.80
	IC	332	296	89.16
Replacement	CA	616	7	1.13
	IC	332	12	3.61

Table 6: Quantitative analysis of **EARAL**’s faithfulness on the ASQA dataset. CA indicates the number of Correct Answers and IC indicates the number of Incorrect Answers. For each scenario, Ablation or Replacement, we analyse queries for which **EARAL**’s initial answer was correct or incorrect. We report the number of analysed queries (# Qu.) and the number of times **EARAL** changes (# Ch.) its answer or attribution with respect to its *Baseline scenario*.

7 Conclusion

In this work, we propose leveraging identifiers of documents in a retrieved context to provide generation-time attribution for LLM-generated answers using a RAG framework. To bridge the gap between model generation and attribution, we propose a two-level attribution method, called **EARAL**, where the document identifier logit is used to estimate token-level contributions, which are then aggregated to compute the statement-level attribution. Extensive experiments using the recent Trust-Align evaluation framework validate the effectiveness, efficiency, and robustness of **EARAL**. Furthermore, regarding faithfulness, our ablation/replacement analysis provides empirical evidence that attribution responds to changes in supporting evidence while remaining stable when irrelevant documents are introduced, a behavior consistent with faithful attribution in settings where the search engine may retrieve irrelevant documents for hard queries.

Limitations

Limitations of our work include investigating only a subset of potential identifier tokens. An interesting extension would be to study newly initialized tokens from the vocabulary, to assess another generalization aspect of our approach. Similarly, while we think the importance is rather the relative order of the target logit values rather than the values themselves during training, this research framework would benefit from more thorough and rigorous analysis concerning the optimal target values. An interesting avenue of research involves extending **EARAL** towards an end-to-end learnable token-level and statement-level attribution. Investigating how our logit-based approach could be

extended to answer token selection and mitigate other types of bias would allow us to enhance, beyond faithfulness, the fairness of model outputs.

Another limitation is related to model size and family of models. Although we explored standard size models (8B), larger models (70B or similar) may be explored but at high computational cost. We limited our comparison to Llama family and we left to future work the exploration of other logits-based models.

Finally, while our faithfulness analysis aimed at demonstrating that **EARAL** showed behavior consistent with faithful attribution, it would be interesting to extend this analysis to a larger-scale with different datasets and comparison to other models.

Ethical Considerations

We acknowledge the ethical risks associated with large language models, including the potential to reproduce or amplify societal biases present in pre-training and fine-tuning data, as well as to generate toxic or misleading content. We did not conduct a systematic audit of fairness or safety in this work and therefore refrain from making claims about bias mitigation or amplification. Residual risks may remain, and a comprehensive assessment is left to future work.

Acknowledgments

This work was carried out as part of a collaboration between IRIT and SNCF, we would like to thank everyone involved for their support and valuable contributions throughout the project. This work was granted access to the HPC resources of IDRIS under the allocation AD011016097R1 made by GENCI.

References

- Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. [QAMPARI: A benchmark for open-domain questions with many answers](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.
- Orlando Ayala and Patrice Bechard. 2024. [Reducing hallucination in structured outputs via retrieval-augmented generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*,

- pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.
- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. 2024. [Contextcite: Attributing model generation to context](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc.
- Xuan-Quy Dao and Ngoc-Bich Le. 2023. Chatgpt is good but bing chat is better for vietnamese students. *arXiv preprint arXiv:2307.08272*.
- Qiang Ding, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2025. [Attention with dependency parsing augmentation for fine-grained attribution](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 372–387, Vienna, Austria. Association for Computational Linguistics.
- Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. [An evaluation framework for attributed information retrieval using large language models](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 5354–5359. ACM.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- 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 reviewing 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.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024a. [Learning fine-grained grounded citations for attributed large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113, Bangkok, Thailand. Association for Computational Linguistics.
- Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024b. [Advancing large language model attribution through self-improving](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3822–3836, Miami, Florida, USA. Association for Computational Linguistics.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).
- Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. 2024. [A peek into token bias: Large language models are not yet genuine reasoners](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4722–4756, Miami, Florida, USA. Association for Computational Linguistics.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. [TinyBERT: Distilling BERT for natural language understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. [A convolutional neural network for modelling sentences](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.
- Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv:2307.16883*.
- Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se-Young Yun. 2021. [Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2628–2635. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. [Hurdles to progress in long-form question answering](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.
- Anson Li, Renee Shrestha, Thinoj Jegatheeswaran, Hannah O Chan, Colin Hong, and Rakesh Joshi. 2024.

- Mitigating hallucinations in large language models: A comparative study of rag-enhanced vs. human-generated medical templates. *medRxiv*, pages 2024–09.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*.
- Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. [ExpertQA: Expert-curated questions and attributed answers](#). 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 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. [Teaching language models to support answers with verified quotes](#). Preprint, arXiv:2203.11147.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. [Peering into the mind of language models: An approach for attribution in contextual question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11481–11495, Bangkok, Thailand. Association for Computational Linguistics.
- Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. [Model internals-based answer attribution for trustworthy retrieval-augmented generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 6037–6053. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Abhilasha Sancheti, Koustava Goswami, and Balaji Srinivasan. 2024. [Post-hoc answer attribution for grounded and trustworthy long document comprehension: Task, insights, and challenges](#). In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 49–57, Mexico City, Mexico. Association for Computational Linguistics.
- Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. [Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Melbourne, Australia. Association for Computational Linguistics.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. [Attribute first, then generate: Locally-attributable grounded text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.
- Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2025. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations*.
- 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.
- Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. [Correctness is not faithfulness in rag attributions](#). In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (IC-TIR)*, ICTIR ’25, page 22–32, New York, NY, USA. Association for Computing Machinery.
- Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. [Unveiling selection biases: Exploring order and token sensitivity in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2025. [Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 969–988, Albuquerque, New Mexico. Association for Computational Linguistics.

Yue Yu, Ting Bai, Hengzhi Lan, Li Qian, Li Peng, Jie Wu, Wei Liu, Jian Luan, and Chuan Shi. 2026. [C2-cite: Contextual-aware citation generation for attributed large language models](#). In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining, WSDM '26*, page 871–880, New York, NY, USA. Association for Computing Machinery.

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

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. [LongCite: Enabling LLMs to generate fine-grained citations in long-context QA](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. [Miracl: A multilingual retrieval dataset covering 18 diverse languages](#). *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *The Twelfth International Conference on Learning Representations*.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. [A setwise approach for effective and highly efficient zero-shot ranking with large language models](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, page 38–47. ACM.

A Datasets Details

A.1 Evaluation Datasets

We follow the *Trust-Align* framework (Song et al., 2025), which builds on ALCE (Gao et al., 2023b)

to assess answer correctness, citation accuracy, and appropriate refusal when retrieved documents lack supporting evidence. ALCE spans three long-form QA datasets: **QAMPARI** (Amouyal et al., 2023) (Wikipedia factoid; answers are sets of entities), **ASQA** (Stelmakh et al., 2022) (ambiguous factoid; multiple concise answers), and **ELI5** (Fan et al., 2019) (open-ended; multi-sentence explanations).

We report Trust-Align metrics on all datasets: **F1AC** (answer correctness on answerable questions), **F1GR** (refusal when unanswerable given retrieved docs), **F1GC** (citation groundedness), and **TRUST** (mean of the three; See (Song et al., 2025) for details). Furthermore, we report the **Average Latency** required to generate and attribute an answer, in ms per query, averaged w.r.t. the three datasets.

A.2 Training Datasets

Training data combines **Trust-Align** (Song et al., 2025), **Hagrid** (Kamalloo et al., 2023), and **ExpertsQA** (Malaviya et al., 2024).

- **Trust-Align**: Builds on QAMPARI, ASQA, and ELI5 and emphasizes robustness to hallucination by including refusal cases when documents lack relevant information (around 25% of training).
- **Hagrid**: Generative Information Retrieval over the MIRACL dataset (Zhang et al., 2023) with human annotations. We keep examples with attributable answer spans and add refusal cases so they comprise 25% of its training set by pairing Hagrid questions with random documents from Hagrid, Trust-Align, and ExpertsQA.
- **ExpertsQA**: Expert-annotated, multi-domain, detailed answers grounded in high-quality sources. We keep attributable spans and similarly add 25% refusal cases by pairing ExpertsQA questions with random documents from the three corpora.

Each example has at least five retrieved documents; if fewer, we pad random documents from the combined pool. We remove bracketed citations from the text. In total, we obtain 20,312 training examples.

B Baselines

We evaluate **EARAL** and zero-shot prompting of Llama 3.1 8B with each marking strategy, and compare against:

- Four baselines reported in (Song et al., 2025): Llama3 8B with *PostCite** (Gao et al., 2023b)

(generate, then retrieve for attribution), *PostAttr** (Gao et al., 2023b) (generate, then entailment via TRUE-NLI), Llama3 8B fine-tuned with *FRONT** (Huang et al., 2024a) (fine-grained citation grounding), and Llama3 8B fine-tuned with *Trust-Align** (Song et al., 2025) (trustworthiness-oriented).

- *ContextCite* (Cohen-Wang et al., 2024) with Llama 3.1 8B (official implementation, default settings), which is a recent *model-internals* approach that learns a surrogate linear model for each query to estimate the contribution the context to the answer.
- *MIRAGE* (Qi et al., 2024) with Llama 3.1 8B (official implementation, default settings), which identifies sensitive tokens and attributes document contributions via input ablations; included as a faithful attribution baseline not previously evaluated on Trust-Align. We furthermore perform significance analysis against this attribution approach as it is the closest to our work, and against the fine-tuned Trust-Align baseline to assess improvements over a strong fine-tuned alternative.
- Llama 3.1 8B: (i) zero-shot with the prompt from (Song et al., 2025); (ii) fine-tuned on the Trust-Align training set for comparison and statistical significance testing.

C Configuration Details

Marking strategy details Identifier tokens are chosen to be (1) single-token under the used tokenizer and (2) extremely rare in training/eval text. After examination, we settled on 10 such tokens as document IDs: {"AA", "BB", "CC", "DD", "EE", "FF", "GG", "HH", "II", "JJ"}. A unique ID is sampled when marking a document.

Attribution aggregation details Attribution aggregates contributions at the sentence level over the generated answer. To enforce trustworthiness, if no document is attributed, we output the *Trust-Align* refusal sentence (Song et al., 2025) as a fail-safe mechanism. For comparability, we use the ALCE top-5 retrieved documents per query. Prompting mirrors *Trust-Align* with a refusal clause; except that we do not request citations.

Identifier token logits learning We set target logit labels ($\hat{l}_k$) empirically, based on training/validation on Hagrid as follows:

- $\hat{l}_k = 4$: for identifier tokens in the ground-truth attribution, $\{id_k \mid id_k \in a_i\}$.
- $\hat{l}_k = 2$: for identifiers in context but not attributed, $\{id_k \mid id_k \in C, id_k \notin a_i\}$.
- $\hat{l}_k = 0$: for identifiers of documents randomly sampled from C .
- $\hat{l}_k = -2$: for identifiers of documents not in C , $\{id_k \mid id_k \notin C\}$.

Hyperparameters settings We train for two epochs with learning rate $2e-5$, a cosine scheduler, and effective batch size 16. For § Eq. 1, we set $\alpha = 0.25$ leveraging preliminary experiments. For aggregation (§ Eq. 5), we use a static attribution threshold $\phi_{\text{prop}} = 3$, selected by optimizing the mean of F1GR and F1GC on the *Hagrid* validation set. We tune $\lambda \in \{0.25, 0.5, 0.75\}$ in a similar manner and use $\lambda = 0.75$ for all results.

All our experiments were performed on a combination of V100s, A100s and H100s GPU cards. A full fine-tuning loop finished in 18 hours on a single A100 GPU.

D Additional Analysis

D.1 Detailed Robustness Analysis

D.1.1 Context Length Analysis

Figure 3 shows our complete robustness analysis regarding the number of documents used in the context. We perform evaluations using the *marking_{BA}* strategy with the top K documents, where K ranges over even integers in $[2, 10]$ plus the default number of documents used in training (i.e., 5).

As expected, the best performing configuration uses 5 documents across all datasets, with performance degrading as we deviate from this value. These findings indicate that optimal performance is closely tied to training configurations, suggesting that more diverse training procedures could improve robustness to varying context lengths.

Nonetheless, **EARAL** exhibits strong robustness to moderate changes in document count. Specifically, TRUST score drops are only 0.63% and 3.10% on ASQA, 5.26% and 3.21% on QAMPARI, and 3.28% and 5.29% on ELI5 when comparing 5 documents to 6 and 4 documents, respectively.

D.1.2 Document Order Analysis

Figure 4 presents comprehensive performance results across different document ordering and iden-

q_1 : Who serves as the president of the senate in the Texas legislature?		
(a ₁) Baseline scenario	(b ₁) Ablation scenario	(c ₁) Replacement scenario
d_1 : < AA>[...] is Texas Lieutenant Governor Dan Patrick [...] d_2 : < BB>[...] Her question was immediately [...] d_3 : < CC>The Lieutenant Governor of [...]	d_1 : < AA>[...] is Texas Lieutenant Governor [removed] [...] d_2 : < BB>[...] Her question was immediately [...] d_3 : < CC>The Lieutenant Governor of [...]	d_1 : < AA>[...] is Texas Lieutenant Governor Dan Patrick [...] d_x : < BB>Mount St. Helens routes can [...] d_3 : < CC>The Lieutenant Governor of [...]
EARAL Answer: The president of the senate, [...], is Texas Lieutenant Governor Dan Patrick.	EARAL Answer: The president of the senate [...] is the Lieutenant Governor of the state of Texas.	EARAL Answer: The president of the senate, [...], is Texas Lieutenant Governor Dan Patrick.
Logits values when generating <i>_Lieutenant</i> : AA = 3.47 , BB = 1.47, CC = 2.71	Logits values when generating <i>_Lieutenant</i> : AA = 2.34, BB = 1.54, CC = 3.08	Logits values when generating <i>_Lieutenant</i> : AA = 3.45 , BB = 0.21, CC = 2.61

q_2 : How many teams in the nba in 1956?		
(a ₂) Baseline scenario	(b ₂) Ablation scenario	(c ₂) Replacement scenario
d_1 : < AA>[...] The NBA had six playoff teams that year [...] d_2 : < BB>[...] the NBA moved to a system in which [...] d_3 : < CC>[...] its smallest size of eight franchises [...]	d_1 : < AA>[...] The NBA had six playoff teams that year [...] d_2 : < BB>[...] the NBA moved to a system in which [...] d_3 : < CC>[...] its smallest size of [removed] [...]	d_1 : < AA>[...] The NBA had six playoff teams that year [...] d_x : < BB>Lady Margaret Butler Lady [...] d_3 : < CC>[...] its smallest size of eight franchises [...]
EARAL Answer: In 1956, the NBA [...], built around eight franchises: [...]	EARAL Answer: I apologize, but I couldn't find an answer to your question [...]	EARAL Answer: In 1956, the NBA [...], built around eight franchises: [...]
Logits values when generating <i>_eight</i> : AA = 1.97, BB = 1.77, CC = 3.71	Logits values when generating <i>_apologize</i> : AA = -0.11, BB = -0.21, CC = 0.04	Logits values when generating <i>_eight</i> : AA = 2.15, BB = 0.08, CC = 3.60

Table 7: Qualitative analysis of faithfulness of **EARAL** on two queries q_1 and q_2 from the ASQA dataset. A comparison of answers and logit value (at a specific generation step) of document identifier tokens is given with respect to three scenarios: Baseline, Ablation, and Replacement. Gold document attribution and associated document identifier token is highlighted in green. Ablated document is highlighted in red. Replaced document is highlighted in yellow.

tifier token selection configurations, evaluated at multiple training checkpoints. We define four experimental configurations:

- **Token selection:** *Rand* (randomly assigned identifier tokens) vs. *Alph* (alphabetically ordered tokens: “AA”, “BB”, etc.)
- **Document ordering:** *Vanilla* (relevance-based ranking) vs. *Rev* (reverse relevance ranking)

We analyze all four combinations (e.g., *Rand-Vanilla*, *Alph-Rev*) to evaluate their joint effects on performance throughout training.

Effect of identifier token selection Comparing *Alph-Vanilla* with *Rand-Vanilla* and *Alph-Rev* with *Rand-Rev* reveals that **EARAL** becomes robust to identifier token choice as early as 40% through the first epoch across all datasets. This rapid adaptation occurs because **EARAL** is trained with random token selection, making alphabetical ordering simply a special case of the training distribution.

Effect of document ordering By comparing *Alph-Vanilla* with *Alph-Rev* and *Rand-Vanilla* with *Rand-Rev*, we observe two key findings: (1) reversed relevance-based ranking slightly improves performance during early training stages, and (2) this advantage diminishes as training progresses,

with all configurations converging in late-stage training. This convergence demonstrates that **EARAL** learns position-invariant document consideration, further evidencing its robustness.

D.2 Proportion Threshold Analysis

The aggregation operator employs hyperparameter λ (§ Eq.5), representing the proportion of statements considered for attribution. While we use $\lambda = 0.75$ for all main experiments (§ C), we analyze the sensitivity of this parameter here.

Table 5 shows TRUST scores for **EARAL** with $marking_{BA}$ across three λ values. Performance remains remarkably stable across all datasets. The TRUST score differences between $\lambda = 0.75$ and $\lambda = 0.25$ are only 2.35% (70.86 vs. 69.19), 3.40% (47.63 vs. 46.01), and 1.63% (48.94 vs. 48.14) for ASQA, QAMPARI, and ELI5, respectively.

This stability demonstrates the confidence of **EARAL**’s debiasing approach. The model assigns consistently high logit values to the same identifier tokens throughout statement generation. Consequently, when identifier logit values are similar at each generation step, varying the threshold proportion from 25% to 75% has minimal impact on the final attribution decisions.

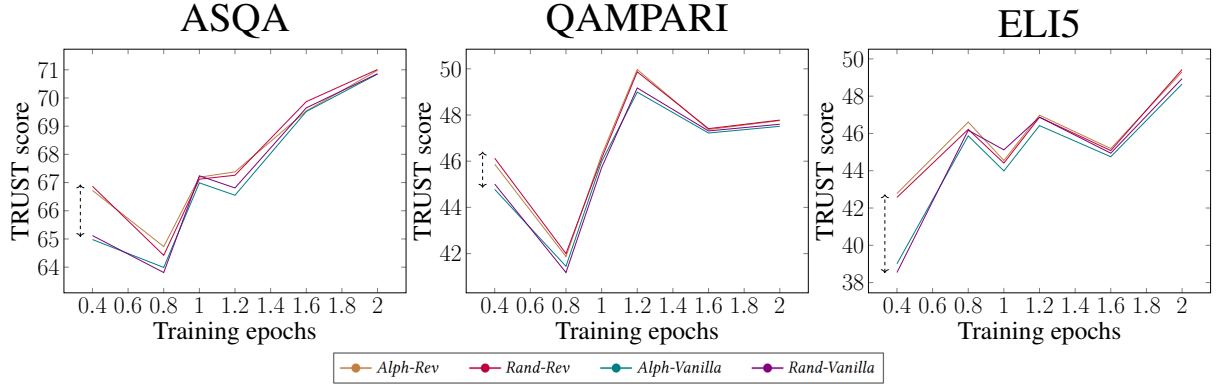


Figure 5: Robustness of **EARAL** to document ordering and identifier token selection. *Rand-Vanilla* (purple) corresponds to the default configuration of our *marking_{BA}* strategy.

D.3 Document order

As mentioned in section 6.3.2, **EARAL** demonstrates robustness to both identifier token selection and document ordering on ASQA, QAMPARI, and ELI5, the three used datasets as shown in Figure 5).

D.4 Qualitative Faithfulness Examples

We conduct a qualitative analysis of **EARAL** on two random answerable examples from ASQA. As discussed in Section 6.4, we consider three scenarios: (a) a *Baseline* RAG setting with gold attribution; (b) an *Ablation* scenario where the answer-bearing span is removed from the gold document; and (c) a *Replacement* scenario where a context document is substituted with an unrelated one. Table 7 presents detailed examples from the faithfulness analysis.

In the *Ablation* cases, removing the relevant span shifts identifier logits away from the original document and leads to either a different answer or abstention, satisfying condition (1). In the *Replacement* cases, the answer remains unchanged and irrelevant documents receive consistently lower identifier logits, satisfying condition (2).