Where Did That Come From? Sentence-Level Error-Tolerant Attribution

Anonymous ACL submission

Abstract

Attribution is the process of identifying which parts of the source support a generated output. While attribution can help users verify content and assess faithfulness, existing task definitions typically exclude unsupported or hallucinated content leaving them unattributed, overlooking the potential to increase faithfulness certainty, locate the error, and fix it easier. In this paper, we propose a new definition for **sentence-level error-tolerant attribution**, which extends attribution to include incorrect or hallucinated content. We introduce a benchmark for this task and evaluate a range of models on it. Our results show that sentence-level error-tolerant attribution improves the quality of both automatic and manual faithfulness evaluations, reducing annotation time by 30% in long-document settings, and facilitates hallucination fixing. We also find that unfaithful outputs are often linked to sentences that appear later in the source or contain non-literal language, pointing to promising avenues for hallucination mitigation. Our approach offers a better user experience along with improved faithfulness evaluation, with better understanding of model behavior.[1]

1 Introduction

Text generation systems are increasingly deployed to produce summaries, answers, and explanations grounded in source documents. A central concern in these applications is *faithfulness*—whether the generated content accurately reflects the input. Unfaithful generations, or hallucinations, can mislead users, damage trust, and propagate misinformation. To address this, recent work has proposed the task of *attribution*: (Bohnet et al., 2022; Gao et al., 2023b; Xu et al., 2025) identifying which parts of the source support a given generation. Attribution can improve trust, let the user expend their knowledge in a certain point, or provide a foundation for verifying outputs.

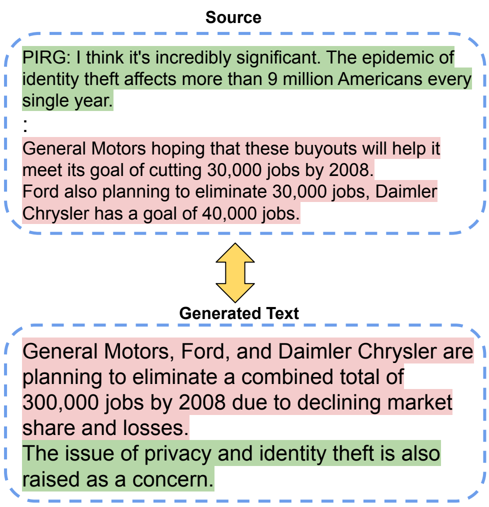


Figure 1: An attribution example adapted from our annotated dataset. The green generated sentence is faithful while the red one is not.

However, existing approaches to attribution face two core limitations. First, most attribution benchmarks operate at the document level (Gao et al., 2023b; Deng et al., 2024), making it difficult and time-consuming to locate the specific source span that is relevant to the output. While some recent work has explored finer-grained attribution at the span level (Huang et al., 2024; Slobodkin et al., 2024), these efforts have largely been limited to the attribute-then-generate paradigm, where relevant source spans are selected prior to generation and used to guide the output. In such setups, attribution is effectively given rather than inferred. These methods do not address the more challenging case where attribution must be extracted retrospectively from existing outputs.

Second, prior work overwhelmingly treats attribution primarily as a form of grounding: if content is not supported by the source, it is simply left unattributed. This framing limits the practical util-

[1]Data will be released upon acceptance.

ity of attribution. Unattributed information can leave users uncertain whether a sentence is entirely fabricated, partially correct, or accurate but misattributed. In contrast, attributing incorrect information can help pinpoint the error, separate accurate content from inaccuracies, and facilitate correction. Such diagnostic benefits are not possible under the current definition of attribution.

To address these two challenges, we propose two key innovations: **sentence-level attribution** in a post-hoc setting, and **error-tolerant attribution**. This sentence-level granularity enables immediate localization of relevant source spans, eliminating the need to read the entire document to verify a single sentence. Building on this, **error-tolerant attribution** extends attribution to cover even incorrect or hallucinated content—whether it contradicts the source, loosely resembles it, or refers to information that arguably should have been included. This richer attribution not only identifies precise source spans (or confirms their absence), but might also offer novel insight into the model's errors and how specific elements in the source may have contributed to them. An example for our new attribution approach is presented in Figure 1.

Our approach provides several benefits across different use cases. First, it serves as a valuable tool for end users on its own. By making fine grained error-tolerant attribution available, it increases the trustworthiness of the output and enables users to verify or expand their understanding without needing to read the entire source document. Moreover, it improves clarity in cases of incorrect output. Users can easily identify and correct errors, or to distinguish between accurate and inaccurate components within a sentence.

Our approach also offers advantages as an auxiliary tool for faithfulness evaluation, whether manual or automatic. By substantially narrowing the scope of information that needs to be considered, it makes it more feasible to annotate or assess long documents with greater consistency and reduced effort.

Finally, this form of attribution serves as a foundation for deeper analysis. It enables tracing the origin of both accurate and hallucinated content, characterizing features that are uniquely associated with hallucinated attributions, and gaining a better nuanced understanding of system behavior.

In this paper, we formalize the task of sentence-level error-tolerant attribution (Section 3) and construct a benchmark to evaluate the performance of various models on it (Section 4, 5). We demonstrate its potential utility for automatic faithfulness evaluation, particularly in challenging scenarios involving long input contexts (Section 6.1). Additionally, we show that our approach reduces manual evaluation time by 30% (Section 6.2). We also show the benefit of using these attribution to fix the output (Section 7). Beyond evaluation and fixing, we analyze the source sentences linked to unfaithful outputs and find that they are more likely to occur toward the end of the document and to contain complex, non-literal expressions (Section 8). These findings highlight promising directions for future work on hallucination mitigation. Overall, our results show that attribution—when extended in this way, becomes a powerful lens for evaluating and improving the faithfulness of text generation systems.

2 Related Work

2.1 Attribution Methods

The attribution task was rigorously defined by Rashkin et al. (2023) as the ability for a generic hearer to say, "According to the *source*, we can infer the *generated-text*," where the source must be interpretable within its context. Three key dimensions characterize existing systems: the overall method type, the granularity of attribution, and when—and how—document retrieval is performed.

Method Type. Attribution methods can be broadly categorized into three paradigms. The *end-to-end* approach generates text alongside citations (Gao et al., 2023b; Deng et al., 2024), while the *post-hoc* approach generates the attribution after the output text already exists (Bohnet et al., 2022; Gao et al., 2023a). More recently, a third paradigm has emerged: *attribute-then-generate*, where relevant spans from source documents are first selected, and then used to condition text generation (Huang et al., 2024; Slobodkin et al., 2024). Although this last approach improves attribution quality in some settings, it is not applicable when the generated text is already fixed. Our work, therefore, focuses on the post-hoc setting, where attribution must be computed retroactively.

Granularity Level. Attribution can vary in granularity on both the output and source sides. On the output side, models have attributed entire responses (Menick et al., 2022; Thoppilan et al., 2022), individual sentences (Gao et al., 2023b; Deng et al., 2024; Slobodkin et al., 2024), or even sub-sentence spans (Xu et al., 2025). On the source side, most systems cite entire documents (Gao et al., 2023b; Deng et al., 2024), primarily due to the difficulty of identifying fine-grained evidence. More recent work has pushed toward concise, localized citations by aligning small spans from the source with specific segments of the generated text (Huang et al., 2024; Slobodkin et al., 2024). However, these methods are limited to the attribute-then-generate paradigm, where attribution is provided prior to generation. In contrast, our work is the first to address fine-grained attribution in the more challenging post-hoc setting, where the generated output is fixed and attribution must be inferred retrospectively. Concurrent work (Zhang et al., 2024) focuses on sentence-level *end-to-end* attribution, however, their approach differs from ours in both setup and objectives, as they do not attribute to incorrect information.

Retrieval Timing. Attribution is often decomposed into two stages: retrieving relevant documents and then identifying evidence within them. For end-to-end and attribute-then-generate methods, retrieval must occur prior to generation, as citations are embedded during text production. In post-hoc setups, retrieval may happen either before or after text generation. Some methods pre-retrieve a document set and restrict attribution to that subset (Bohnet et al., 2022), while others generate text freely and then retrieve supporting evidence afterward (Gao et al., 2023a). A third, less common variant is the "closed-book" approach, which avoids retrieval entirely and relies solely on the model's internal knowledge. This approach consistently underperforms and is typically used as a baseline (Bohnet et al., 2022; Gao et al., 2023b). In our work, we assume a fixed set of input documents, representing an early retrieval step and focusing the problem on evidence selection rather than retrieval.

Overall, in the aforementioned prior work, attribution has been treated as evidence-based: text lacking faithful grounding should not receive attribution. In contrast, we expand this definition to include even unfaithful generations. This extension enables new uses: attribution can serve as a means of verifying faithfulness, localizing hallucinations, and potentially correcting them. Although Gao et al. (2023a) did attempt to connect generated text with source content post-hoc to fix its output, their work did not isolate attribution as a standalone task, nor was attribution quality evaluated independently. Our work introduces and formalizes this task, demonstrating its value in analyzing and improving faithfulness in generation.

Generated Sentence (g): *The mayor introduced a new climate initiative earlier this week.*	
Category	**Source Sentence** ($s_i \in \mathcal{D}$)
Evidence	The mayor announced a new climate initiative on Monday.
Contradiction	The mayor explicitly denied any plans for a new climate initiative.
Near Match	The mayor announced a new recycling program on Monday.
Expected Span	The mayor supported the budget. [No climate content; mayor mentioned once.]
None	[No climate or mayor content; or- frequent mayor mentions.]

Table 1: Examples of attribution categories given a single generated sentence.

2.2 Reference - Source Alignment

Another related line of work focuses on aligning spans in reference summaries with source documents. Such alignments have been used to automatically generate training data for summarization subtasks such as salience detection (Gehrmann et al., 2018; Lebanoff et al., 2019), redundancy elimination (Cho et al., 2019), and text fusion (Zhang et al., 2018; Lebanoff et al., 2019). More recently, alignment itself has been framed as an independent task (Ernst et al., 2021), enabling the creation of more accurate alignment datasets and models (Ernst et al., 2024), which in turn can enhance end-to-end summarization systems (Ernst et al., 2022). Our work draws inspiration from this alignment perspective, but shifts focus to the generated text, which may include hallucinations and inaccuracies—posing a fundamentally different challenge than aligning human-authored summaries.

3 Task Definition

Let $\mathcal{D} = \{s_1, s_2, \ldots, s_n\}$ denote the set of **source sentences** from a document or collection of documents, and let g be a **generated sentence** produced by a system grounded in $\mathcal{D}$. The goal is to identify a minimal subset $A \subseteq \mathcal{D}$ that *maximizes the information relevant to assessing the faithfulness* of g to the source. The selected subset A must be interpretable in context to resolve any coreferences.

Each attribution set A may fall into one or more of the following categories (An example for each category can be found in Table 1):

Evidence. Sentences that directly support or entail the content in g, indicating that it is faithful.

Contradiction. Sentences that explicitly contradict information in g, suggesting unfaithfulness.

Near Match. Sentences that closely resemble g with small, non-contradictory variations.

Expected Span. Sentences (or regions of the document) where the information in g would reasonably be expected to appear if it were grounded in the source. The absence of relevant content in such expected spans may imply unfaithfulness.

None. No sentence in $\mathcal{D}$ is relevant (or all are equally relevant); g is likely a hallucination.

In the ideal case of perfect attribution, the selected set A^* should be sufficient to annotate the faithfulness of g. More precisely, once A^* is known, the remainder of the document provides no additional information about g in the context of assessing its truthfulness or provenance. Formally, this implies $I(g\,; \mathcal{D} \setminus A^* \mid A^*) = 0$ where $I(\cdot; \cdot \mid \cdot)$ denotes conditional mutual information. Therefore, if A^* falls under the *Near Match* or *Expected Span* categories, then under the assumption of perfect attribution, g can be labeled as unfaithful without requiring access to the full document.

4 Data Annotation

In order to evaluate baselines for this task and present the potential of optimal attributions in different scenarios, we annotate manually a development and test attribution sets.

4.1 Dataset

We leverage TofuEval (Tang et al., 2024), a recent benchmark that comprises two summarization datasets: MediaSum (Zhu et al., 2021), which summarizes dialogues, and MeetingBank (Hu et al., 2023), which summarizes meeting transcripts.

TofuEval sampled 50 documents from each dataset. For every document, three topic titles were generated, and six different LLM-based summarization models (OpenAI's GPT-3.5-Turbo, Vicuna-7B (Chiang et al., 2023) and WizardLM7B/13B/30B (Xu et al., 2024), and one anonymized model) produced a summary focused on each topic. Each sentence in these summaries was then manually annotated for faithfulness to the source with a binary label, an error type, and a detailed explanation of the error.

This setup yields a total of 2 datasets $\times$ 50 documents $\times$ 3 topics $\times$ 6 systems $=$ 1800 summaries. The dataset is split into a development set and a test set, containing 70 and 30 documents, respectively. Due to the high annotation cost, we randomly selected one system-generated summary per topic, resulting in $2 \times 50 \times 3 \times 1 = 300$ summaries annotated, while preserving the original development and test split.

Since we also aim to identify attributions for incorrect summary sentences, we require generated summaries that contain hallucinations, preferably accompanied by detailed explanations. TofuEval is, to the best of our knowledge, the only dataset that includes detailed faithfulness annotations (including explanations) for relatively long documents (averaging 950 words) and captures real-world errors produced by recent LLM-based models—making it ideally suited for our purposes.

4.2 Annotation Process

Our annotation task aims to identify alignments between source document sentences and corresponding summary statements to serve as attributions. We followed the annotation protocol of Slobodkin et al. (2022), using controlled crowdsourcing (Roit et al., 2020). Potential annotators were prescreened through a filtering task and underwent multiple training phases of increasing difficulty to ensure quality. Ultimately, 10 qualified annotators completed the task.

We employed the web-based annotation tool from Slobodkin et al. (2022), deployed via Mechanical Turk[2] (see Figure 3 in the Appendix). The interface presents the document and summary side by side. Annotators were instructed to focus on one summary sentence at a time, aiming to align the entire summary by the end of the task.

To facilitate annotation, annotators were told to

[2] www.mturk.com

select standalone sub-sentence spans from the summary and match them to full sentences from the source document. This encouraged them to distinguish between accurate and inaccurate portions of the summary sentence and to find appropriate source sentences for both. We later aggregated these annotations to the sentence level, mapping each summary sentence to a set of document sentences. To streamline the task further, when an incorrect summary sentence was selected, we displayed the corresponding error explanation from TofuEval in the annotation interface. Full annotation guidelines are provided in Appendix A.

4.3 Data Quality

To assess the quality of alignments we measured inter-annotator agreement in pairs on a set of instances annotated by all annotators, comparing 430 pairs of annotations. For each pair, we computed the intersection-over-union (*IoU*) of token indices (restricted to content words) in the document spans aligned to the same summary sentence, following Ernst et al. (2021). The resulting average *IoU* was 0.47, indicating moderate agreement.

A manual inspection revealed that while annotators consistently identified the core attribution sentences, the precise sentence set boundaries were often ambiguous. This subjectivity explains the moderate agreement score.

To validate annotation quality, an expert evaluated a subset of 10 documents, 30 summaries, and totaling 83 summary sentences. For each summary sentence, the expert assessed the correctness of linked document sentences (Precision). We also measured Recall against an expert-level gold attribution set, which consists of the original annotated attributions supplemented with additional source sentences identified by an expert to ensure full coverage. The average Precision was 94.37%, and Recall was 90.27%, indicating high data quality.

5 Experiments

To establish the performance of current large language models (LLMs) on the task of post-hoc fine-grained attribution, we evaluate several state-of-the-art models. The objective is to assess their capability in identifying precise spans within source documents that support or relate to specific segments of a given generated text, as defined in Section 3 and annotated according to Section 4. We experiment with zero-shot setting the following LLMs: *Gemini 2.0 Flash*, *GPT-4o* (Hurst et al., 2024), *Qwen 2.5 72B Instruct* (Team, 2024) and *Llama 3 70B Instruct* (Grattafiori et al., 2024).

Performance is measured using sentence-level macro-averaged Precision (P), Recall (R), and F1-Score (F1). These metrics compare the model's predicted attribution spans against the human-annotated gold standard from our test set.

Model	Macro P	Macro R	Macro F1
Qwen 2.5 72B	0.5230	0.6419	0.5360
GPT-4o	0.5672	0.6474	0.5723
Llama 3.3 70B	0.5738	0.6452	0.5640
Gemini 2.0 Flash	0.6032	0.6939	**0.6075**

Table 2: Performance of LLM baselines on the fine-grained attribution task.

Table 2 presents the performance of the LLM baselines. The results indicate that current general-purpose LLMs can perform this fine-grained attribution task to a notable extent. Gemini 2.0 Flash in the zero-shot setting achieves the highest performance across all metrics, with a Macro F1-Score of 0.6075. Llama 3.3 70B (zero-shot) and GPT-4o (zero-shot) also demonstrate competitive results, outperforming Qwen 2.5 72B.

Upon analyzing the errors of the best-performing model, Gemini, we found that it reliably identifies the core attribution sentences in most cases. For simpler, more extractive summary sentences, this leads to highly accurate results. However, in more subjective cases, Gemini tends to include additional source sentences that are already conceptually covered by the core attributions. As a result, while the model's output is generally of high quality, it is still not always as concise or tightly scoped as desired.

6 Faithfulness Evaluation

Beyond the benefits of error-tolerant attribution as a standalone tool, we also aim to demonstrate its value as an auxiliary task across several scenarios. In this section, we show how it can support and enhance both automatic and manual faithfulness evaluation. In Section 7, we illustrate how it can aid in correcting hallucinations. Finally, in Section 8, we show that our annotated dataset reveals features that may help identify, in advance, text spans that are more likely to produce hallucinations.

		Summac-zs	Summac-conv	AlignScore	Llama-3.1	Vicuna-1.5	Mistral	Gemma-3	Claude-3.5	GPT-4o
1 doc	Plain	55.95	57.75	**71.58**	59.88	50.27	57.29	70.94	69.91	**75.72**
	Highlighted	N/A	N/A	N/A	63.00	52.92	57.03	**71.47**	69.37	74.37
	Attr. Only	**60.06**	**60.33**	65.00	**67.96**	**64.19**	**59.47**	66.30	**72.20**	67.62
	Attr. Only + Incor. Orig.	42.60	63.51	68.96	71.93	34.82	50.74	59.16	70.62	66.03
10 docs	Plain	43.05	50.00	66.93	57.51	N/A	49.88	60.67	62.81	71.18
	Highlighted	N/A	N/A	N/A	60.55	N/A	50.07	59.92	72.94	70.75
	Attr. Only	60.06	60.33	65.00	67.96	N/A	59.47	**70.96**	72.20	67.62
	Highlighted Gemini	N/A	N/A	N/A	64.48	N/A	54.09	65.42	67.64	72.29
	Attr. Only Gemini	**63.22**	**62.16**	**67.30**	**68.69**	N/A	**61.31**	68.37	**73.37**	**73.60**

Table 3: Performance of different models in faithfulness evaluation with original source, with highlighted attribution, or with attribution only.

6.1 Assisting Automatic Evaluation

We evaluated the ability of different models to assess the faithfulness of summary sentences, both with and without attribution. Given a source document and a generated summary sentence, the task is to classify whether the sentence is faithful to the source. We explored three input formats for providing the source: (1) the full source document ('Plain'), (2) the source with highlighted sentences that are attributed to the summary sentence ('Highlighted'), and (3) only the attributed sentences, presented without context ('Attr. Only').

We compared a range of models over the entire test set, including non-LLM-based factuality metrics (SummaC-ZS, SummaC-CV (Laban et al., 2022), and AlignScore (Zha et al., 2023)), open-source LLMs (Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Vicuna-7B-v1.5 (Chiang et al., 2023), Mistral-8B-Instruct (Jiang et al., 2023), and Gemma3-4B-it (Team et al., 2025)), and proprietary LLMs (GPT-4o (Hurst et al., 2024) and Claude-3.5-haiku). For evaluation, we used balanced accuracy (Laban et al., 2022; Tang et al., 2023), which accounts for label imbalance between faithful and unfaithful cases.

As shown in Table 3, most models performed better when provided with highlighted or attributed content, with attribution-only mode often yielding the highest accuracy, albeit sometimes by a small margin. To further demonstrate stronger benefits of attribution, we examined long context where each document was randomly shuffled into a group with nine other randomly selected documents from the development set. Under this condition, the attribution-only mode consistently outperformed others, with larger performance gains, likely due to the reduced distraction from irrelevant context.

We also applied the same 10-document analysis using the predicted attributions from the best-performing model in Section 5, Gemini. Surprisingly, Gemini's attribution-only setup outperformed most other configurations—including the gold attribution-only setup. This suggests that LLM-generated attributions may be better aligned with how models interpret and utilize information, compared to human-annotated ones.

To examine the role of error-tolerant attribution, we compared the standard attribution-only setup (using a single document) to a mixed setup in which only faithful summary sentences are evaluated with attribution, while unfaithful sentences are evaluated using the full source without attribution. This simulates a case where error-based attributions are unavailable. As seen in the results, most models exhibited performance degradation in this mixed setup, sometimes substantially. In the few cases where performance improved, the gains were small and might be due to response variability. These findings highlight that error-based attributions typically enhance model evaluation performance, or at the very least, do not harm it.

6.2 Assisting Manual Evaluation

Manual evaluation is both costly and time-consuming, particularly for long documents. To ease this burden, we investigate whether sentence-level attribution can assist annotators in evaluating both correct and incorrect summary sentences.

To that end, we recruited four NLP research students as expert annotators to assess the faithfulness of summary sentences with respect to the source documents. Each annotator was assigned documents in one of two modes: (1) a plain document with no attribution, or (2) an interactive version where clicking a summary sentence highlighted its attributed sentences in yellow. Annotators could use Ctrl+F to search for keywords in both modes. To mitigate bias, each document was evaluated only once by a given annotator and in only one mode. Each document was annotated twice, once per condition, by different annotators. In total, we randomly selected one summary per document for all 30 test set documents, resulting in 76 summary sentences that were evaluated. Similar to the automatic evaluation (Sec. 6.1), we used the balanced accuracy to aggregate the annotations.

Our results show that attribution reduced annotation time by 30% (76s vs. 108s per summary sentence) while also slightly improving balanced accuracy (82.05% vs. 77.90%)[3]. All annotators reported that the highlights were helpful, indicating the potential of sentence-level attribution to support manual evaluation. We hypothesize that this benefit would be even greater in settings involving longer documents or more abstractive summaries, where keyword search is less effective.

Notably, there was no significant difference in annotation time between correct and incorrect summary sentences. This suggests that, in the absence of attribution, incorrect sentences may require even more time to evaluate, further emphasizing the value of error-tolerant attribution.

We also tested a similar setup on 14 development-set documents, annotated by two annotators using predicted highlights from the best model in Section 5, Gemini. While annotation time was similarly reduced by 28% (86s vs. 118s), attribution performance dropped with highlights (68% vs. 77%). Annotators noted that the predicted highlights were less focused than the gold ones, often spanning multiple paragraphs and making them harder to follow, though sometimes helpful by showing repeated information. This observation aligns with our error analysis in Section 5.

7 Fixing Unfaithful Text

In this section, we explore another potential application of error-tolerant attribution: correcting unfaithful text. While post-editing techniques aim to resolve inconsistencies in generated content (Dong et al., 2020; Balachandran et al., 2022; Gao et al., 2023a), they remain challenging due to the difficulty of localizing errors and identifying the correct information. To address this, we propose using error-tolerant attribution to guide the system's attention toward the relevant source content, thereby facilitating more effective correction.

Given an unfaithful summary sentence and its source, the goal is to minimally revise the sentence so that it becomes faithful to the source. We evaluate three input configurations: (1) the full source without attribution information ('Plain'), (2) the source with highlighted attribution spans ('Highlighted'), and (3) only the attribution spans without additional context ('Attr. Only'). We generated corrections of 20 incorrect summary sentences, using three models—LLaMA 3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-8B-Instruct (Jiang et al., 2023), and Claude-3.5-Haiku. The prompts are presented in Appendix D. An expert annotator then assessed each corrected sentence with two judgments: whether the sentence was fixed, partially fixed, or not fixed, and its relative ranking across the three input settings for each model.[4]

	Input	Fix Status			Ranking		
		F	PF	NF	1	2	3
Mistral	Original	5%	5%	90%	45%	30%	25%
	Highlighted	5%	10%	85%	40%	45%	15%
	Attr. Only	15%	5%	80%	70%	25%	5%
LLaMA	Plain	15%	20%	65%	20%	40%	40%
	Highlighted	45%	5%	50%	50%	45%	5%
	Attr. Only	65%	5%	30%	60%	25%	15%
Claude	Plain	45%	10%	45%	50%	35%	15%
	Highlighted	50%	30%	20%	55%	20%	25%
	Attr. Only	30%	30%	40%	15%	50%	35%

Table 4: Fix Status (Fixed/Partially Fixed/Not Fixed) and Ranking (1 is best) Percentages by Model and Input Format

As shown in Table 4, for all three models, the inclusion of highlighted attributions led to better corrections and higher rankings compared to the plain source. The attribution-only setting performed best with Mistral and LLaMA, but was least effective with Claude. This suggests that Claude benefits more from contextual grounding, while the smaller models gain more from direct, focused attribution cues. Overall, these results highlight the utility of using sentence-level, error-tolerant attributions to guide factual corrections—particularly for smaller models, where attribution spans help isolate relevant content and reduce the cognitive and computational burden of processing entire documents.

8 Analysis of Hallucination Factors

The centrality of the hallucination problem in text generation raises a fundamental question: *what*

[3]Four summary sentences with ambiguous faithfulness labels and explanations were excluded from this analysis.

[4]Rank percentages may sum to more or less than 100% due to tied scores.

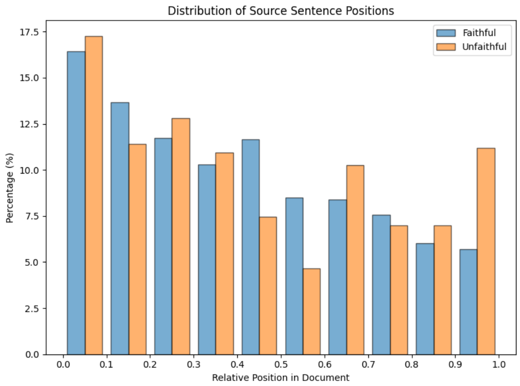


Figure 2: Comparing position of faithful-linked sentences to unfaithful-linked sentences.

causes models to hallucinate? Our manually annotated dataset offers fertile ground for such an investigation. By analyzing source-based features that may contribute to unfaithful outputs, such as ambiguity or complexity, we can identify characteristics that predispose certain input content to hallucinations. This opens up the possibility of pre-editing sources (e.g., simplifying complex segments) to reduce hallucination risk, or at least to anticipate it. Prior work has explored various contributing factors, including the influence of the prompt (Rawte et al., 2023; Yao et al., 2023), the training process (Li et al., 2024), and the training data (Dziri et al., 2022). However, to the best of our knowledge, no study has examined the source of hallucinations in relation to the *input* text itself. Our approach aligns with the goals of Koniaev et al. (2025), who focused on identifying problematic sources that tend to yield less informative summaries.

To conduct our analysis, we used the source sentences that are manually linked to summary sentences in our dataset. These sentences are implicitly selected by the model for generating a summary. We divided them into two groups: those linked to at least one *unfaithful* summary sentence, denoted as ‘unfaithful-linked sentences’, and those linked only to *faithful* ones, denoted as ‘faithful-linked sentences’. We then examined several features to identify signals that could distinguish between the two groups, and found two particularly informative ones: sentence position and the presence of non-literal expressions.

Sentence Position. We computed the relative position of each source sentence in its document and plotted the distributions (Figure 2). As expected—and consistent with prior findings in summarization research (Lebanoff et al., 2019)—faithful-linked source sentences are more likely to appear at the beginning of the document, with their frequency gradually decreasing as the document progresses. Interestingly, unfaithful-linked sentences exhibit a different distribution: aside from a slight peak in the first 10% of the document, their occurrence is more evenly spread across positions. Notably, in the final 10% of the document, the likelihood of a sentence being linked to unfaithful content is nearly double that of faithful-linked sentences. This suggests that models may struggle to accurately process or incorporate information from later parts of the input.

Non-Literal Language. We also examined the prevalence of non-literal expressions (e.g., idioms, irony), which require additional interpretation or external knowledge. A manual review of 10 development-set documents (covering 30 summaries and 84 summary sentences) revealed that 25% of the sentence sets linked to unfaithful summary sentences contained at least one non-literal expression—compared to only 9% among those linked to faithful sentences. Viewed from another perspective, 88% of the non-literal expressions that are linked to the summary, led to unfaithful sentences. These findings suggest that non-literal language, which is inherently harder to interpret, increases the likelihood of unfaithful generation. Examples can be found in Appendix E.

In sum, our analysis underscores the value of investigating hallucination through the lens of *source-based features*. While our analysis is exploratory, it highlights promising directions for future work aiming to discover and leverage additional features to combat hallucination.

9 Conclusion

We introduced a fine-grained, error-tolerant approach to attribution that operates post-hoc at the sentence level, enabling both accurate localization of source evidence and meaningful interpretation of unfaithful outputs. Our benchmark demonstrates the utility of this framework for faithfulness evaluation, significantly reducing annotation effort and providing deeper insight into model behavior. By extending attribution beyond faithful outputs, we show its potential as both a practical tool for users and a diagnostic signal for improving text generation systems.

Limitations

Our findings highlight the utility of sentence-level, error-tolerant attribution across several use cases. However, our conclusions are based on experiments with only two datasets, both from the domain of dialogue summarization. These were the only available datasets that met our criteria: recent model outputs, existing faithfulness annotations, high rates of hallucinations, and sufficiently long input texts. As a result, the generalizability of our conclusions to other tasks, such as question answering, remains uncertain and warrants further investigation.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

Haolin Deng, Chang Wang, Li Xin, Dezhang Yuan, Junlang Zhan, Tian Zhou, Jin Ma, Jun Gao, and Ruifeng Xu. 2024. WebCiteS: Attributed query-focused summarization on Chinese web search results with citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15095–15114, Bangkok, Thailand. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. Proposition-level clustering for multi-document summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1765–1779, Seattle, United States. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Aviv Slobodkin, Sharon Adar, Mohit Bansal, Jacob Goldberger, Ran Levy, and Ido Dagan. 2024. The power of summary-source alignments. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6527–6548, 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.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. 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. 2024. 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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Steven Koniaev, Ori Ernst, and Jackie Chi Kit Cheung. 2025. Presumm: Predicting summarization performance without summarizing. *arXiv preprint arXiv:2504.05420*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Vipula Rawte, Prachi Priya, SM Tonmoy, SM Zaman, Amit Sheth, and Amitava Das. 2023. Exploring the relationship between llm hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *arXiv preprint arXiv:2309.11064*.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled Crowdsourcing for High-Quality QA-SRL Annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. 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.

Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst, and Ido Dagan. 2022. Controlled Text Reduction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5699–5715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. 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 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. ALiiCE: Evaluating positional fine-grained citation generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 545–561, Albuquerque, New Mexico. Association for Computational Linguistics.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

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

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. 2024. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Appendix

A Dataset Creation

A.1 License

TofuEval dataset that serves as the basis to our benchmark, is released with MIT license, and is allowed for academic purposes.

A.2 Attribution Annotation Guidelines

Definition: Attributed Source Sentences: Attributed source sentences help a human faithfulness verifier assess whether a summary span is faithful to the source text.

An attributed source sentence may serve as:

- **Evidence**: Directly supports the summary sentence.
- **Contradiction**: Directly contradicts the summary sentence.
- **Close Paraphrase (but not identical)**: Contains similar information with slight modifications (e.g., "Ori went to the beach" instead of "Aviv went to the beach").
- **Contextual Anchor**: A sentence where we would expect the information to appear if it were explicitly mentioned.

Matching Guidelines

Breaking the summary sentence into propositions

The worker should break down the summary sentence into standalone (non-consecutive) parts (propositions). Usually, each part contains a main verb.

Example: John went home and ate an apple

- John went home
- John. . . ate an apple

Rules of thumb

- As a rule of thumb - each standalone verb should be in a different proposition.

Example:

The Federal Reserve is expected to continue raising interest rates to cool down the economy , which has been experiencing a slowdown .

- The Federal Reserve is expected to continue raising interest rates to cool down the economy
- the economy , which has been experiencing a slowdown .

The buyouts , negotiated with the United Auto Workers Union , will provide lump sum payments of up to $140,000 .

- The buyouts , negotiated with the United Auto Workers Union
- The buyouts…will provide lump sum payments of up to $140,000

Example with additional verbs:

The document notes that the U.S. government has stated that Iraq has no weapons of mass destruction , which is a lie , and that the U.S. is not going to wait for countries like Iraq declared to be part of the so - called axis of evil to develop weapons of mass destruction .

- The document notes that the U.S. government has stated that Iraq has no weapons of mass destruction , which is a lie
- and that the U.S. is not going to wait for countries like Iraq declared to be part of the so - called axis of evil to develop weapons of mass destruction .
- Rule of thumb - if the document sentences that align with a single summary sentence are not consecutive and each document sentence corresponds to a different part of the summary sentence, then those parts of the summary sentence should also be separated.
- Rule of thumb - Try to separate the supported and unsupported parts of the summary sentence, if each part can standalone and be separated. (even if the document sentences are consecutive)
- In many cases, the summary sentence contains only a single proposition.
- In general, propositions that separating them would change the meaning significantly, like in the case of reason and cause, may not be separated in some cases.
- Both sides (reason and cause) can be part of the span

"John ate an apple due to his hunger."

- To decide - this rule can be applied for reason and cause as well. - Rule of thumb - if the sentences in the document that align with a single summary sentence are not consecutive and each document sentence corresponds to a different part of the summary sentence, then those parts of the summary sentence should also be separated.

The matching described below should be done from a summary span (proposition) to a set of document sentences.

Alignment Boundaries

- Match a summary proposition to document full sentences.
- When highlighting from the document side, assume we have the context of this sentence. Therefore, no need to assign another sentence just for the name of the speaker (for instance). We know it as we have context.
- Be concise. Only if a single document sentence does not cover the summary proposition in full, add more document sentences.

Supported/Unsupported labeling

- For each summary sentence, the worker gets a former annotation of whether this sentence is supported by the document or not, and an explanation why not.
- This information should help the workers in their annotation.
- However, if the worker disagrees with the former annotation, they are allowed to change it.
- Additionally, for an "unsupported" sentence, the "unsupporting" label may not apply to all spans within the sentence. In such cases,

the annotator should update the label to "supported" for spans that are supported, while retaining the "unsupporting" label only for spans that lack support.

- The "unsupported" explanation can lead the worker where to look for the mistake. Even if there are several options for where the mistake comes from, choose the one that is mentioned in the explanation.

- The "unsupported" explanation can help the worker to break the summary sentence into pieces (propositions), as in many cases the explanation focuses on one part that is not supported where the rest is supported, or two different unsupported parts.

Select the strongest evidence available

- If an exact supporting/unsupporting sentence exists, do not select weaker alternatives (e.g., a close paraphrase).

- Select only the strongest evidence (or closest sentence)

- If multiple sentences provide equivalent evidence, match all of them separately.

Ensure full coverage of the summary sentence

- The summary sentence should be covered in full.

- Breaking the summary sentence into standalone pieces should help you to assure each part is aligned properly.

Handling Missing or Implicit Information

- If a piece of information is not explicitly mentioned in the text and there is no closely related sentence that could be a corrupted version,

- In some rare cases, the topic of this piece of information is mentioned only in a single sentence or paragraph. In these rare cases, you can align this sentence or paragraph, as the information would be expected to appear if it were present in the text.

- In most cases, where the topic is related to many areas from the document, and it is not directly connected to a specific paragraph, the attribution is None.

A.3 Annotation Interface

Figure 3 presents a printscreen of the annotation interfaces used during the crowdsourcing. Annotators were paid 13$ per hour with additional bonuses awarded for high-quality work.

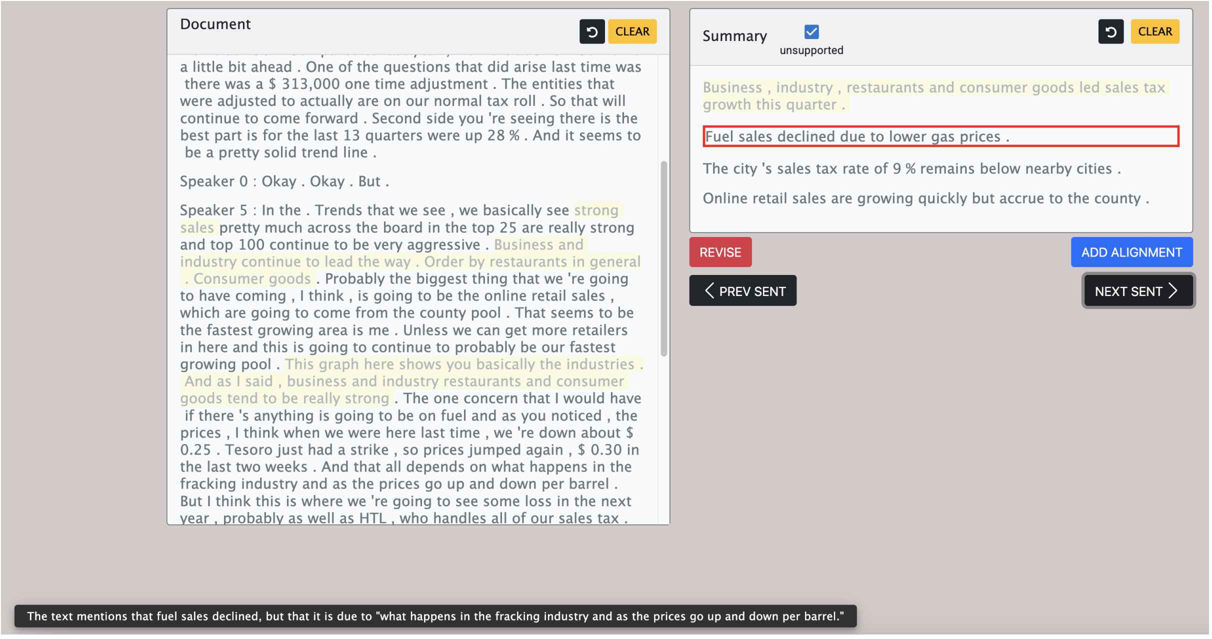


Figure 3: The alignment annotation interface. The annotator marks a span (proposition) in the summary (right) along with all matching spans in the current document (left). To minimize cognitive load, visual focus is placed on one summary sentence at a time (red rectangle) to orient the process. Additionally, by hovering over the "supported" checkbox, whenever a reason for unsupportedness is provided by the original annotation, it is presented to the annotators (black textbox) to help in the annotation process.

B Automatic Faithfulness Evaluation

We used a single RTX8000 to run all the evaluation experiments. For long documents some models required more computation resources, so we used a cluster of 4 GPUs. On average, it took around 1 hour per model. The prompt we used for all LLM-based models can be found in Table 5.

C Manual Faithfulness Evaluation

C.1 Technical Details

In this experiment, we measured both quality and work time. To that end, we added a timer that starts automatically when the annotator reveals a new document by an additional clicking, and not earlier when accepting the task. The timer can be paused if the annotator needs a break. This improves the less accurate previous approach (Akoury et al., 2020; Krishna et al., 2023)that measures the time by the difference between task submission times.

C.2 Expert Training

We designed a training task consisting of three summaries—two with highlighted source sentences and one without. Only annotators who performed well on this task, achieving high accuracy against gold labels, were selected to continue with the annotation process. Ultimately, we hired four expert annotators, all of whom are AI research students. They were compensated at a rate of $25 per hour.

C.3 Annotation Guidelines

C.3.1 Not Highlighted Task

For each summary sentence, decide if it is faithful to the document. Don't be stressed by the timer. Take the time in order to make the correct decision. There are no highlights in this part. If it helps, you can use Ctrl+F to look for relevant keywords.

If you are not sure, select faithful. Ignore small nuanced shifts between the summary and the source.

C.3.2 Highlighted Task

Don't be stressed by the timer. Take the time in order to make the correct decision. Click a summary sentence to highlight the most relevant document sentences for it. These highlights should help you to make your decision and should be sufficient in most cases. The rest of the document is provided for context. The highlights may include contradictions or instances where some information from the summary sentence is absent in the source. These cases should be marked as unfaithful. If the highlights are not enough, you can use Ctrl+F to look for keywords. If you are not sure, select faithful.

Attribution Task Annotation Prompt

You are an **expert annotator** performing an **attribution task**. Your goal is to identify the source sentences within a document that are most relevant to assessing the faithfulness of a given summary sentence.
Task Definition: Given a summary sentence and a document (presented as a list of indexed sentences), find the "attribution" for the summary sentence within the document.
Attribution Definition: Attribution is defined as a *minimal set* of document sentences that maximally supports the certainty of a reader in assessing the faithfulness of the summary sentence. This means finding the fewest document sentences that contain the core information needed to judge if the summary sentence is accurate, contradictory, or closely related to the document's content. The attribution could be:

- Evidence supporting the summary sentence.
- Sentences contradicting the summary sentence.
- Sentences containing very similar text or concepts, but not exactly the same.
- Sentences indicating the location where the information *should* logically be found, even if it slightly differs.
- If the summary sentence appears entirely fabricated or has no plausible basis in the document, the attribution is `None`.

Input:
Summary Sentence:

```
{summarySentence}
```

Document Sentences (with indices):

```
{list_of_indexed_document_sentences}
```

Instructions:

1. Read the Summary Sentence carefully.
2. Read through the Document Sentences.
3. Identify the sentence indices from the Document Sentences that form the minimal attribution set according to the definition provided.
4. Focus on the *most essential* sentences needed to verify or contradict the summary's claim.
5. If no relevant sentences are found (summary is fabricated relative to the document), output `None`.

Output Format: Output *only* a Python list containing the integer indices of the identified document sentences. For example: `[18]` or `[5, 6]` or `[21, 23]`. If the attribution is `None`, output the word `None`. Do not include any explanations or additional text.
Output:

Figure 4: Prompt used to guide models in identifying source attribution

Ignore small nuanced shifts between the summary and the source.

C.4 Annotation UI

We used Mechanical Turk Sandbox platform (free of charge) in order to provide the annotators an accessible format. A snapshot is shown in Figure

Setup	Prompt
System Prompt	`You are a helpful assistant evaluating factual consistency between a summary sentence and a source text.` `Given the source and the summary, answer with 'yes' if the summary is faithful to the source, or 'no' if it is not.`
Plain	`Is the evaluated summary sentence faithful to the source?  Reply only with ýesór ńo.́`
Highlighted	`The source text includes special [FOCUS]...[/FOCUS] tags marking parts that are the most relevant source sentences to the evaluated summary sentence. Is the evaluated summary sentence faithful to the source?  Please use the marked source sentences to help you decide.  Reply only with 'yes' or 'no'.`
Attribution Only	`The following Relevant Source Sentences were extracted from the source as the most relevant information in the source to the evaluated summary sentence.` `Based on the Relevant Source Sentences alone, is the evaluated summary sentence faithful?  Reply only with 'yes' or 'no'.`

Table 5: Prompt used for automatic faithfulness evaluation.

5.

D Fix Hallucination Evaluation

All evaluation experiments were conducted on a single RTX8000 GPU. The prompt used for GPT-4o is provided in Table 6.

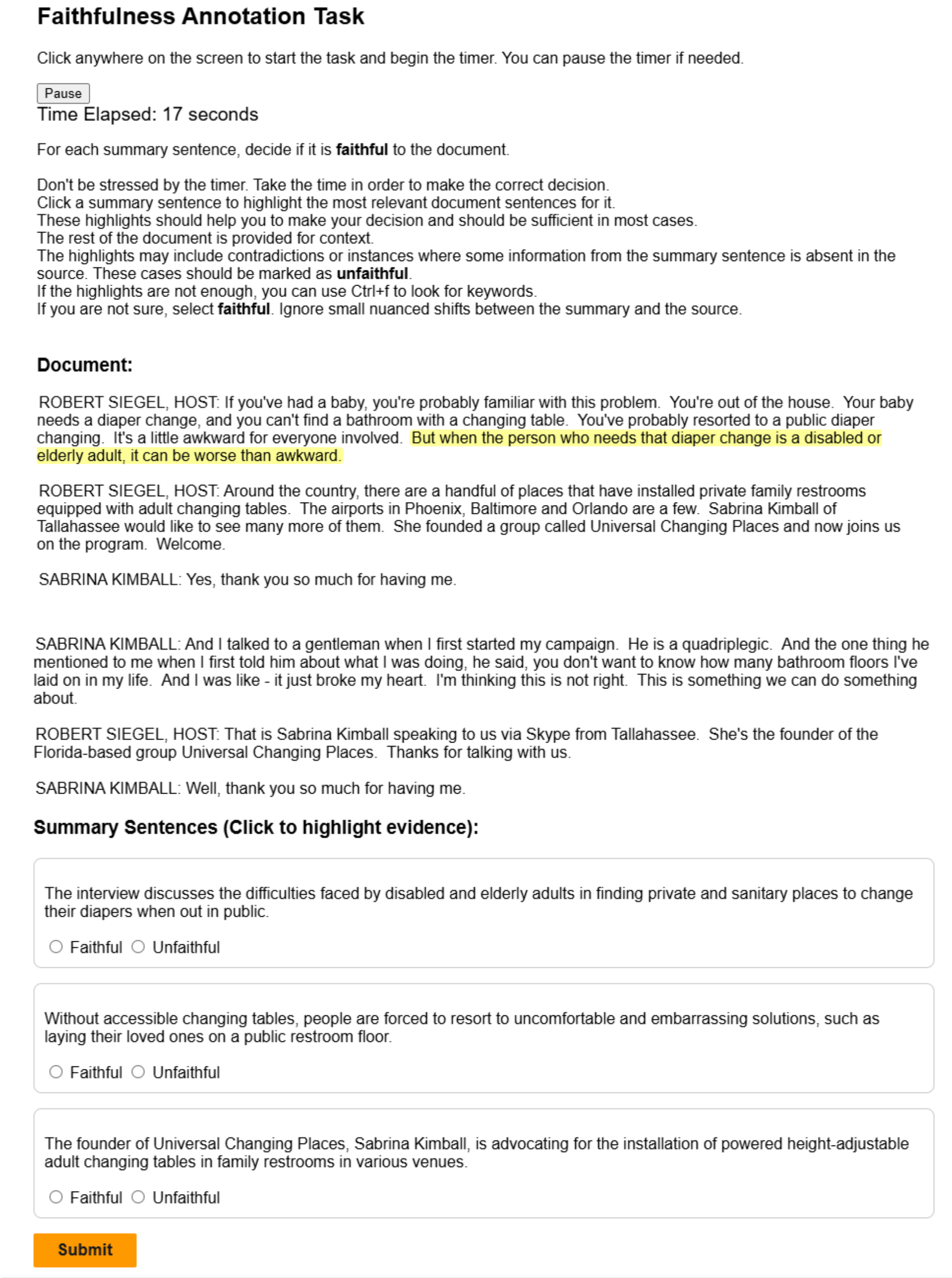
E Non-Literal Expression Examples

Here are some examples for non-literal expression we have found in the source that are linked to an unfaithful summary sentence.

- This was *a perfect storm of disaster* that actually probably saved his life because when the airplane ascends, you lose oxygen, the air gets thin as we would say in layman's terms.
- *Net net*, it was about a $2,000 loss, which sounds like a lot of money, but it was a million and a half dollars worth of bonds.
- And also Fidel Castro is *on his last legs*, so to speak.
- We saw very different answers depending on who in Congress was asking him the question, but I think the overall takeaway point here is that he got trapped when he wanted *to put his foot down* and have strong answers and show the President that he wasn't going to be bullied by Congress, then he had something to say.
- The Internal Revenue Service saying that we would be *sharing our personal tax information to protect the privacy of our tax information.*

F Use Of Ai Assistants

We have used AI to improve writing, mostly for paraphrasing, and also to facilitate coding in certain parts. We went over all code/text paraphrased or generated by AI and verified its correctness.


Faithfulness Annotation Task

Click anywhere on the screen to start the task and begin the timer. You can pause the timer if needed.

Pause

Time Elapsed: 17 seconds

For each summary sentence, decide if it is **faithful** to the document.

Don't be stressed by the timer. Take the time in order to make the correct decision.
Click a summary sentence to highlight the most relevant document sentences for it.
These highlights should help you to make your decision and should be sufficient in most cases.
The rest of the document is provided for context.
The highlights may include contradictions or instances where some information from the summary sentence is absent in the source. These cases should be marked as **unfaithful**.
If the highlights are not enough, you can use Ctrl+f to look for keywords.
If you are not sure, select **faithful**. Ignore small nuanced shifts between the summary and the source.

Document:

ROBERT SIEGEL, HOST: If you've had a baby, you're probably familiar with this problem. You're out of the house. Your baby needs a diaper change, and you can't find a bathroom with a changing table. You've probably resorted to a public diaper changing. It's a little awkward for everyone involved. But when the person who needs that diaper change is a disabled or elderly adult, it can be worse than awkward.

ROBERT SIEGEL, HOST: Around the country, there are a handful of places that have installed private family restrooms equipped with adult changing tables. The airports in Phoenix, Baltimore and Orlando are a few. Sabrina Kimball of Tallahassee would like to see many more of them. She founded a group called Universal Changing Places and now joins us on the program. Welcome.

SABRINA KIMBALL: Yes, thank you so much for having me.

SABRINA KIMBALL: And I talked to a gentleman when I first started my campaign. He is a quadriplegic. And the one thing he mentioned to me when I first told him about what I was doing, he said, you don't want to know how many bathroom floors I've laid on in my life. And I was like - it just broke my heart. I'm thinking this is not right. This is something we can do something about.

ROBERT SIEGEL, HOST: That is Sabrina Kimball speaking to us via Skype from Tallahassee. She's the founder of the Florida-based group Universal Changing Places. Thanks for talking with us.

SABRINA KIMBALL: Well, thank you so much for having me.

Summary Sentences (Click to highlight evidence):

The interview discusses the difficulties faced by disabled and elderly adults in finding private and sanitary places to change their diapers when out in public.

○ Faithful ○ Unfaithful

Without accessible changing tables, people are forced to resort to uncomfortable and embarrassing solutions, such as laying their loved ones on a public restroom floor.

○ Faithful ○ Unfaithful

The founder of Universal Changing Places, Sabrina Kimball, is advocating for the installation of powered height-adjustable adult changing tables in family restrooms in various venues.

○ Faithful ○ Unfaithful

Submit

Figure 5: The manual faithfulness evaluation interface. The document is exposed and the timer begins only after reading the instructions and clicking the screen. The timer can be paused manually. We present here only an excerpt of the full document.

Setup	Prompt

System Prompt	`You are a helpful assistant fixing summary sentences to be faithful to the source.`
	`Given the source and an unfaithful summary sentence, fix the summary sentence with minimum changes so it will be faithful to the source.`
	`Write only the fixed sentence without any additional text or explanation.`

Plain	`Fix the summary sentence to be faithful with minimum changes.`

Highlighted	`The source text includes special [FOCUS]...[/FOCUS] tags marking parts that are the most relevant source sentences to the evaluated summary sentence.`
	`Based on the Relevant Source Sentences alone, fix the summary sentence to be faithful with minimum changes.`
	`Please use the marked source sentences to help you decide.`

Attribution Only	`The following Relevant Source Sentences were extracted from the source as the most relevant information in the source to the evaluated summary sentence.` `Based on the Relevant Source Sentences alone, fix the summary sentence to be faithful with minimum changes.`

Table 6: Prompt used for automatic hallucination fixing