Attribute First, then Generate:
Locally-attributable Grounded Text Generation

Aviv Slobodkin[1*]   **Eran Hirsch**[1*]   **Arie Cattan**[1]   **Tal Schuster**[2]   **Ido Dagan**[1]

[1]Bar-Ilan University   [2]Google Research

{lovodkin93, hirscheran, ariecattan}@gmail.com
talschuster@google.com   dagan@cs.biu.ac.il

Abstract

Recent efforts to address hallucinations in Large Language Models (LLMs) have focused on attributed text generation, which supplements generated texts with citations of supporting sources for post-generation fact-checking and corrections. Yet, these citations often point to entire documents or paragraphs, burdening users with extensive verification work. In this paper, we introduce a *locally*-attributable text generation approach, prioritizing concise attributions. Our method, named *"Attribute First, then Generate"*, breaks down the conventional end-to-end generation process into three intuitive steps: content selection, sentence planning, and sequential sentence generation. By initially identifying relevant source segments (*"select first"*) and then conditioning the generation process on them (*"then generate"*), we ensure these segments also act as the output's fine-grained attributions (*"select"* becomes *"attribute"*). Tested on Multi-document Summarization and Long-form Question-answering, our method not only yields more concise citations than the baselines but also maintains—and in some cases enhances—both generation quality and attribution accuracy. Furthermore, it significantly reduces the time required for fact verification by human assessors.[1]

1 Introduction

Grounded text generation, which includes tasks like summarization and question-answering, aims to produce content that is derived from specific sources, either user-provided or retrieved via retrieval mechanisms (e.g., RAG; Lewis et al., 2020). To facilitate verification of models' adherence to these sources, recent years have seen a growing interest in *attributed* text generation (Bohnet et al.,

2023), which aims to create text alongside supporting evidence. This method enhances models' credibility by enabling factuality verification. It also aids in detecting and addressing factual errors, a critical need given the observed frequency of such errors, termed "hallucinations", in model outputs compared to source texts (Mishra et al., 2024a).

While these attributions are aimed to facilitate factuality evaluation by focusing people's attention on relevant supporting texts, current approaches often yield rather *coarse* attributions, pointing back to whole documents or paragraphs. Such attributions, though better than having none, require human assessors to exhaustively sift through many irrelevant details in the cited content, resulting in a time-consuming and somewhat ineffective fact-checking process. Can we do better?

In this work, we extend attributed text generation to also consider attribution conciseness, reformalizing the task as *Locally-attributable Text Generation* (see Section 2). We introduce a criterion for attributions to be *precise*, targeting only the most relevant text snippets, from specific sentences down to sub-sentence spans, while avoiding non-essential details. This criterion complements the *full coverage* requirement, which stipulates that every generated fact be backed by a cited snippet. As illustrated in Fig. 1, given a query and several documents, the model output includes citations to relevant spans within the source documents for each generated sentence instead of broadly attributing the entire documents (Gao et al., 2023b). These spans represent the minimal set of relevant source snippets needed to cover the content of the output sentences.

Following our localization criterion, we propose a novel attribution-driven generation approach, aimed to ensure both *conciseness* and *full coverage*, termed *"Attribute First, then Generate"*. Rather than jointly producing text with citations (Gao et al., 2023b) or getting the attribution post-generation (Bohnet et al., 2023), we
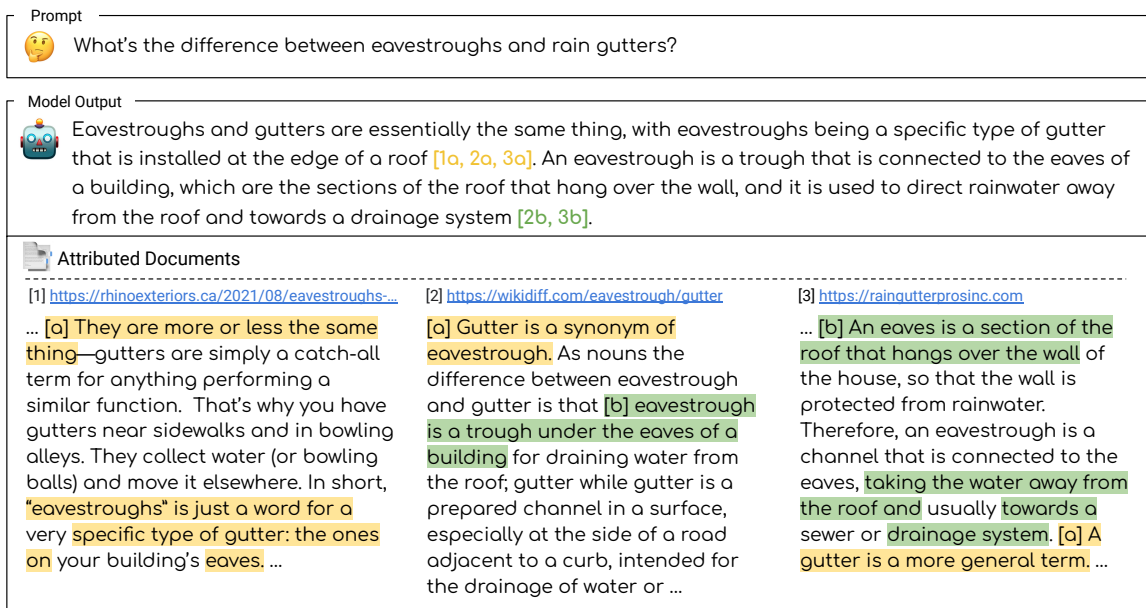

Figure 1: In our approach, demonstrated here on LFQA, the input to the model assumes a question and a collection of documents, which could be retrieved or provided by the user. The generated text is grounded in these documents, accompanied with sentence-level citations to concise relevant spans from the supporting sources. Previous approaches (e.g. Gao et al., 2023a,b) cite the whole source (truncated here for brevity), making it harder for readers to reliably verify the generation.

propose to *first* choose relevant details from the source text(s), and *then* focus the entire generation process on them. Specifically, we decompose the conventional end-to-end grounded generation approach into three intuitive steps, as depicted in Figure 2: (1) content selection (choosing relevant details from source texts), (2) sentence-level planning (organizing and grouping content for sentence fusion), and (3) sentence-by-sentence generation (based on the selected and structured content). Notably, by dictating the generation process, the initially selected details de facto serve as attributions. This ensures a closer attribution adherence to the generated output, as the entire generation process is now guided by these attributions. It also offers more precise supporting evidence, in the form of these selected details, thereby fulfilling our objectives of attribution accuracy and conciseness.

Our method, adaptable to various grounded generation tasks, requires only slight adaptations in the content selection step to cater to specific lexical needs, e.g., salience in summarization or query relevance in question-answering. We showcase its effectiveness in both Multi-document Summarization (MDS) and Long-form Question-answering (LFQA), exploring both finetuning and few-shot prompting strategies. Evaluations across automatic and human assessments reveal that our strategy not

only meets but in some cases surpasses existing baselines in generation quality while also achieving high attribution accuracy. Importantly, it also leads to significantly shorter citations, thereby reducing the effort needed for fact-checking by over 50%.

Altogether, our work is the first to propose localized attributions and to develop a method geared to provide such concise citations. This level of granularity in attribution enhances the efficiency of the fact-verification process, by focusing users' attention on the most relevant details. We thus suggest that future research will build upon our paradigm to further advance locally-attributed generation.

2 Locally-attributable Grounded Text Generation

The task of Locally-attributable Grounded Text Generation involves processing a given set of documents $\mathcal{D}$ to produce an output text y. The set of documents may either be provided by the output of a retrieval system or by the user. The output, consisting of sentences $s_1, \ldots, s_n$, should both fulfill the base underlying task (e.g., summarizing for MDS or generating a response for LFQA) and also include attributions $C_1, \ldots, C_n$ pointing back to supporting text from $\mathcal{D}$. For these attributions to effectively facilitate factuality verification by human
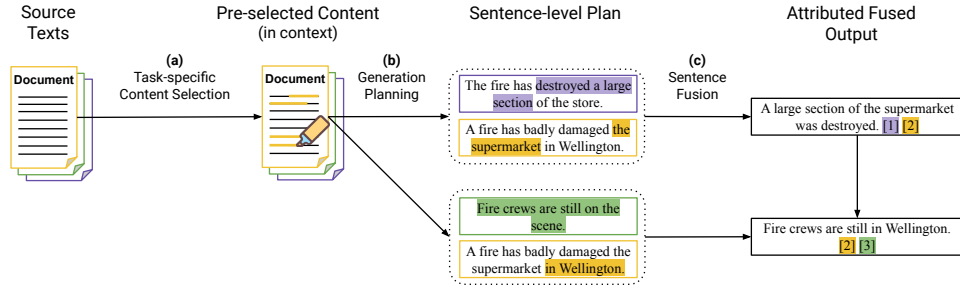
Figure 2: Our *attribute first* process guides the model to output fluent texts that are consistent with input sources, and include fine-grained sentence-level attributions to localized text spans (i.e., highlights) from the inputs.

readers, it is vital that they support as much of the output as possible, while also achieving the highest granularity in both the input and output.

Concretely, *output* granularity entails assigning the smallest information units in the generated text y—in this work, individual sentences s_i—with a corresponding attribution set C_i, consisting of one or more supporting citations $c_i^j \in C_i$. Simultaneously, *input* granularity requires each citation c_i^j to point to a specific span of text $h_k \in \mathcal{D}$, possibly non-consecutive. Importantly, these spans should be as concise as possible while ensuring completeness in attribution, i.e., that every piece of information in the output, which in our setting refers to each output sentence s_i, is fully backed by the union of its citations C_i. For example, in Fig. 1, C_i is depicted as squared brackets at the end of each sentence with pointers to spans h_k, shown as colored highlights in $\mathcal{D}$ (we use "highlights" and "spans" interchangeably).

3 Modeling

This work aims to integrate source attribution with the content-grounded generation process, to achieve more targeted attribution. Our approach is based on uncovering the inherent decision-making mechanisms involved in text generation. We first describe our suggested *Attribute First, then Generate* scheme in Section 3.1. Then, we propose two strategies for this framework's application: a prompt-based in-context learning approach (Section 3.2) and fine-tuning designated components (Section 3.3). For each approach, we provide a brief overview of our implementation strategy for the framework's phases, while a more in-depth elaboration can be found in Appendix C.

3.1 Attribute First, then Generate

To reliably achieve the desirable fine-grained attribution, we introduce the *Attribute First, then Generate* paradigm, outlined in Fig. 2. This method is structured around three key steps, mirroring the inherent decision-making processes involved in text generation: content selection, sentence-planning, and sentence-by-sentence generation. Though applicable to any grounded text generation task, we specifically test its application in Multi-document Summarization (MDS) and Long-form Question-answering (LFQA), adapting only the content selection step to each task's unique requirements.

Content Selection The approach begins with identifying relevant spans from the source that would contribute information to the generated output ((a) in Fig. 2). These spans then function as a more focused grounding content, guiding the lexical choices in the generation process. Notably, as these chosen segments constitute the primary information for the resulting text, they effectively serve as its (fine-grained) attribution.

Sentence Planning This step is designed to achieve sentence-level attribution in the *output* through the introduction of an intermediate sentence-level planning step ((b) in Fig. 2). Here, the selected spans from the previous step are organized into coherently ordered clusters, where the spans grouped within each cluster naturally fit within the same sentence. This step, which breaks down the constrained generation process into smaller, more manageable steps, serves two purposes: it simplifies the model's task by conditioning each generation iteration on a single cluster rather than the entire set of selected spans, while also leading to the desired, sentence-level attribution in the *output*, in the form of the corresponding cluster's highlights.

3311

3

Sentence-by-sentence Generation In the final step ((c) in Fig. 2), the model executes the text generation according to the outlined plan, adopting a sentence-by-sentence approach to preserve the built-in attribution from the previous steps. Specifically, it uses the next set of highlights C_{i+1} and all preceding sentences $s_{1:i}$, to generate the next sentence s_{i+1}, by maximizing $p(s_{i+1}|s_{1:i}, C_{i+1})$. This ensures each sentence expresses the content of the designated highlights while also integrating seamlessly with the preceding output.

3.2 In-context Learning

For each of the substeps delineated in the previous section, we propose a prompt-based in-context learning (ICL) strategy, where a model is guided with specific instructions and carefully selected few-shot examples. The content selection step ((a) in Fig. 2) is formulated as $\mathcal{D} \rightarrow h_1, \ldots, h_m$, where $\mathcal{D}$ is the set of source documents and h_i corresponds to a verbatim span from $\mathcal{D}$. Namely, a model is tasked with outputting relevant spans from $\mathcal{D}$, separated by a designated markup (see prompts in Figs. B.1 and B.5). In MDS, these spans are aimed to capture salient information, while for LFQA, the model, presented with a query, is instructed to identify spans answering that query. These spans are then located within the source documents via string-matching, with any undetectable span being omitted (for more details, see Appendix C.1). Notably, in all subsequent steps, we "highlight" the identified spans within the document set $\mathcal{D}$ via designated markups before and after each span, akin to the methodology used in prior work (Gehrmann et al., 2018; Slobodkin et al., 2022, 2024).

For the sentence planning step ((b) in Fig. 2), the documents, now "highlighted", are concatenated, truncating each at the final highlighted span to manage input length and focus the model's attention. To highlight, we use two designated markups, *<highlight_start>* and *<highlight_end>*, which we insert into the text before and after each consecutive 'highlighted' span, respectively (see prompts in Figs. B.2 and B.6). The model is instructed to organize $h_1, \ldots, h_m$ into an ordered set of highlight clusters $C_1, \ldots, C_n$, where $C_i = \{h_k\}_{k=1}^{|C_i|}$, to guide the subsequent generation phase. For more details, see Appendix C.2.

Finally, in step (c), the model is guided to generate the output one sentence at a time, adhering to the structured plan. Specifically, we separate

the generation process into a series of model calls, each focusing on a single cluster's highlights, concatenated in the same manner as in step (b). The model is also presented with all previously generated sentences $s_{1:i}$. The task is then structured as maximizing $p(s_{i+1}|s_{1:i}, C_{i+1})$, guiding the model to produce the subsequent sentence s_{i+1}, considering both the current cluster of highlights C_{i+1} and the previously-generated sentences $s_{1:i}$.[2] For more details, see Appendix C.3 as well as the full prompts structure in Figs. B.3 and B.7.

Chain-of-Thought. We also explore a variation that combines the sentence-planning and sentence-by-sentence generation stages ((b) and (c)), inspired by the Chain-of-Thought paradigm (CoT; Wei et al., 2022). Given the selected content from step (a), the model is tasked with generating sentence-level plans, deciding on highlights to include in each synthesized sentence, alongside producing these sentences (see prompts in Figs. B.4 and B.8). The model is guided to produce the final output, after forming the plan, focusing on unifying and refining sentences for coherence. For sentence-level attribution, we link each sentence in the output with the highlights from its planning stage in the CoT process.[3] For more details, refer to Appendix C.3.

3.3 Fine-tuning

We also explore fine-tuning models for the various steps of our framework, aligning the input-output formulation of each subtask to its ICL counterpart in Section 3.2. During content selection ((a) in Fig. 2), our model is taught to copy task-specific segments from $\mathcal{D}$ verbatim, and to separate them using a special markup (see Appendix C.1 and Figs. B.9 and B.12). To further ensure strict lexical fidelity to the source text, a constrained decoding approach is adopted during inference (see Appendix C.6 for more details).

In the sentence planning step ((b) in Fig. 2), we use special tokens to 'highlight' the previously selected spans, and teach our model to copy verbatim the "highlighted" spans in a clustered and ordered manner, using special tokens to separate between different clusters and different highlights within a cluster (see Appendix C.2 and Figs. B.10 and B.13 for more details). This phase similarly incorporates

---

[2] In Section 5.1 we show that conditioning on $s_{1:i}$ improves coherence.

[3] We use spaCy sentencizer (Honnibal and Montani, 2017).

a constrained decoding technique to maintain exact lexical matching to the highlighted content. Finally, for the sentence generation step ((c) in Fig. 2), we use special tokens to separate between the highlighted documents, the prefix, and for the LFQA setting the query (see Appendix C.3 and Figs. B.11 and B.14).

Joint Selection and Planning. We also explore a joint strategy that combines the content selection and plan generation stages ((a) and (b)). Here, the entire document set $\mathcal{D}$ serves as input, and the output is an ordered series of "highlight" clusters $C_1, \ldots, C_n$. To ensure exact copying of source spans, we employ the aforementioned constrained decoding strategy. For more details, see Appendix C.2.

4 Experimental Setup

4.1 Datasets

To construct the training, development, and test splits for the different components within our framework, we employ source-output alignment datasets that map corresponding spans between inputs and outputs. These alignments are adapted to suit the data requirements of each framework phase: for content selection, the alignments' *source*-side spans serve as the target labels; in sentence planning, source spans corresponding to the same output sentence are clustered, with the order of output sentences determining the order of these clusters; for the sentence-by-sentence generation phase, the output sentences act as the target outputs for their respective source clusters.

For MDS, we use an alignment dataset derived from the DUC, TAC benchmarks[4] and the Multi-News dataset (Fabbri et al., 2019), comprising of pairs of summaries, corresponding source articles and summary-source alignments from Ernst et al. (2022, 2024). For LFQA, we rely on human-annotated alignments of large language model responses with citations (Liu et al., 2023). By filtering responses that were rated as verifiable, we ensure high-quality alignment for this setting. Further details and dataset processing information are provided in Appendix A.

4.2 Models and Baselines

Our study investigates both ICL prompt-based LLMs and fine-tuned LMs.

In-context Learning. For our ICL approach (Section 3.2), we utilize the Gemini model (Gemini Team Google, 2023), applying designated few-shot prompts tailored to each subtask.[5] For each setting, we also implement two baseline models for comparative analysis. For the first baseline, we prompt Gemini to perform the overarching task, namely summarization for MDS and answering a query for LFQA, end-to-end (see Appendix C.4). This baseline is primarily used to ensure the generated text's quality is not compromised in our framework. For the second baseline, we implement ALCE (Gao et al., 2023b), a prompt-based approach designed to enable the simultaneous generation of text and its attributions through in-line citations (e.g., [1] to cite source 1). See Appendix C.5 for more details.

Fine-tuned Models. For each of our proposed subtasks, we fine-tune models on our derived datasets (Section 3.3). We choose to use PRIMERA (Xiao et al., 2022), a Longformer Encoder-Decoder model (LED; Beltagy et al., 2020) optimized for MDS, for both of our MDS and LFQA settings. Our choice of model stems from PRIMERA's leading performance in the MDS setting, attributed to its capability to handle long contexts and its pre-training technique. Further, we found that PRIMERA outperformed the Long-T5 large model (Long-T5$_{\text{Large}}$; Guo et al., 2022) on our LFQA development set, where both models contain roughly the same number of parameters (Long-T5: 750M, PRIMERA: 447M). Analogous to our ICL approach, we also fine-tune a variant to perform the corresponding task comprehensively for both settings.

4.3 Evaluation

Following our task definition in Section 2, we aim to assess two aspects of the produced outputs. Firstly, we evaluate the generated texts' quality, focusing on coherence and task fulfillment—specifically, summary quality in MDS tasks and answer relevance in LFQA, consistent with standard evaluation practices in each respective task (Fabbri et al., 2021b; Stelmakh et al., 2022). Secondly, we examine the attribution quality, looking at how well each sentence is supported by its accompanying citations and the conciseness of these citations (simply measured by the length of the cited texts). Our evaluation methodology combines automated metrics and human judgments.

---

[4]https://duc.nist.gov/

[5]We used the Gemini API, which is free for academic purposes: https://ai.google.dev/models/gemini

	Method	ROUGE$_L$↑	BERTSCORE↑	AUTOAIS↑	LENGTH↓	NO ATT. (%)↓
ICL	GEMINI	19.0	**86.1**	-	-	-
	ALCE	**20.3**	**86.1**	**88.7**	843.6	3.4
	ATTR. FIRST	16.7	85.2	79.5	92.7	0.0
	ATTR. FIRST$_{CoT}$	19.5	**86.1**	72.8	75.3	1.5
FT	PRIMERA	19.9	85.8	-	-	-
	ATTR. FIRST	16.5	84.9	64.4	**33.1**	0.0
	ATTR. FIRST$_{joint}$	18.0	85.3	48.9	57.1	0.0

Table 1: MDS results of both in-context learning (ICL) and fine-tuned (FT) models. The left two columns are metrics that measure the quality of the output, ignoring attributions. The right columns report the attribution quality, including the length (number of tokens) of the cited text. NO ATT. shows the percentage of generated sentences lacking any attribution.

	Method	ROUGE$_L$↑	BERTSCORE↑	AUTOAIS↑	LENGTH↓	NO ATT. (%)↓
ICL	GEMINI	33.1	89.6	-	-	-
	ALCE	35.2	89.9	49.8	2153.3	26.9
	ATTR. FIRST	35.8	90.5	78.7	65.2	0.0
	ATTR. FIRST$_{CoT}$	**38.6**	**90.7**	**89.3**	48.2	0.0
FT	PRIMERA	32.2	88.8	-	-	-
	ATTR. FIRST	24.7	88.0	52.7	**20.9**	0.0
	ATTR. FIRST$_{joint}$	27.0	88.1	44.9	33.1	0.0

Table 2: LFQA results of both in-context learning (ICL) and fine-tuned (FT) models. The left two columns are metrics that measure the quality of the output, ignoring attributions. The right columns report the attribution quality, including the length (number of tokens) of the cited text. NO ATT. shows the percentage of generated sentences lacking any attribution.

Automatic Evaluation. To assess compliance with each task, we employ the ROUGE$_L$ (Lin, 2004) and BERTSCORE (Zhang et al., 2020) metrics. These compare generated texts with reference outputs, namely gold summaries for MDS and reference answers for LFQA.

For assessing citation conciseness, we report the average length (number of tokens) of the cited content per output sentence. To evaluate citation quality, we employ AUTOAIS, an NLI-based metric for attribution evaluation, in line with prior studies (Bohnet et al., 2023; Gao et al., 2023b) which have shown its correlation with human-judgment of attribution quality. Following those works, we employ the TRUE model from Honovich et al. (2022) for our NLI framework.[6] The final AUTOAIS score is the average number of sentences predicted as attributed by the classifier.

Lastly, we also report the average number of non-attributed sentences per instance.[7]

Human Evaluation. The evaluation of generated attributed texts consists of two steps. First, annotators are asked to evaluate text coherence on a 5-point Likert scale, and in the LFQA setting also its helpfulness in addressing the query. Next, annotators proceed to evaluate the quality of citations. We follow the sentence-level attribution evaluation protocol from Gao et al. (2023a), which extends the *Attributable to Identified Sources* (AIS; Rashkin et al., 2022) human evaluation framework of attribution to the sentence-level. Annotators judge whether each output sentence is fully supported by its attribution, scoring 1 for full support and 0 otherwise. The final AIS score is the average of these scores across sentences. For more details, see Appendix G.

5 Results and Analyses

We provide results of our automated evaluation in Tables 1 and 2 and human evaluation in Tables 3 and 4. Considering the resource-intensive aspects of human evaluation, we limit this assessment to the ALCE baseline and our paradigm's most efficient model, the ICL variant ATTR. FIRST$_{CoT}$, chosen for its optimal balance between task performance and attribution accuracy, as indicated by the automatic evaluation.

We first note that our approach, which extracts focused spans from the input documents, leads to

---

[6]We also examined a more recent NLI model (TrueTeacher; Gekhman et al., 2023), and have found TRUE to correlate better with human judgment. See Appendix B for more details.

[7]Given the absence of reliable automatic metrics for fluency, we restrict its evaluation to the human evaluation, following standard practices (Fabbri et al., 2021a).

Method	Fluency↑	AIS↑	Time (sec)↓
ALCE	4.7	**89.3**	47
ATTR. FIRST$_{ICL-CoT}$	**4.9**	79.9	**22**

Table 3: Human evaluation on the MDS outputs. Fluency evaluates the text coherence. AIS assesses the text's attribution support for each sentence, with full support scoring 1, and 0 otherwise. Time is a measurement of the amount of seconds it took an annotator to verify if a text is fully supported by its attribution.

significantly shorter attributions than the ALCE baseline, where citations coarsely point to entire documents. For example, the ATTR. FIRST$_{CoT}$ variant's citations are on average 45 times shorter than ALCE's.[8] Importantly, we find that assessing our paradigm's attribution takes annotators roughly half the time compared to those of ALCE, emphasizing the benefit of attribution localization in streamlining the manual fact-checking process.

In terms of attribution accuracy, our method surpasses the ALCE baseline in the LFQA domain (both on the automatic AUTOAIS and the human AIS metrics), while in MDS, although ALCE shows higher accuracy, it leads to twice as many *un*attributed sentences. We also note that while in the MDS setting our method may decrease attribution accuracy compared to ALCE's full-document citations, readers can always fall back to examining the full document if the provided highlights are insufficiently supporting. Indeed, analyses of the human annotations revealed that around 42% of the 109 sentences deemed unsupported were partially supported, with the remainder of the attributing details frequently found in immediate proximity to the highlighted sections (usually within the same or adjacent sentences).[9] This observation indicates that while our sentence fusion step occasionally includes slightly more information than the selected highlights (later used as fine attributions), it mostly adds details from the surrounding context.[10] Hence, while suggesting room for improvement in strict adherence to highlights, this does not significantly hinder fact-checking effectiveness.

Importantly, the promising results on attribution do not compromise the quality of outputs. Our automatic analyses indicate minimal impact on

Method	Fluency↑	Helpful.↑	AIS↑	Time (sec)↓
ALCE	**4.9**	**4.7**	87.6	59
ATTR. FIRST$_{ICL-CoT}$	**4.9**	4.5	**94.4**	**35**

Table 4: Human evaluation on the LFQA outputs. Metrics are similar to Table 3. The helpfulness metric measures the helpfulness of responses to queries on a 5-point Likert scale.
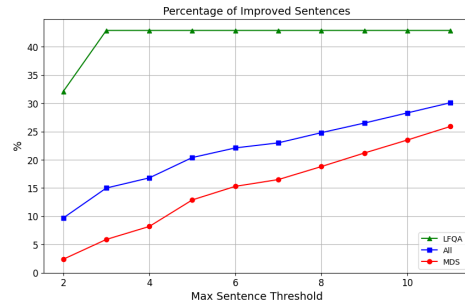

Figure 3: Percentage of more cohesive sentences when including the prefix in the input across varying upper limits of sentence index in the output.

summary quality in MDS and even an improvement in LFQA, as confirmed by the ROUGE$_L$ and BERTSCORE metrics. These results are further evident in our human evaluation, both in terms of the output's fluency and (for LFQA) in helpfulness for answering the query. Notably, our in-context variants tend to produce higher-quality outputs than fine-tuned models. This apparent advantage, observed also in attribution accuracy, is likely due to Gemini's stronger ability to generalize to long-context settings, as well as its larger model size.

Overall, our experiments show that our approach leads to orders of magnitude shorter attributions, with little impact on the attribution accuracy and the output's quality. We leave further improvements in fine-tuning the individual designated components to future research.

5.1 Do Models Benefit from Seeing the Prefix?

We also examine whether providing models with the context of previously generated sentences indeed aids the generation process, particularly in enhancing output cohesion as defined by Maimon and Tsarfaty (2023). Cohesion involves sentences being *"referentially linked"*, e.g., via co-referring mentions, or linked through discourse connectors. To analyze the impact of the prefix on cohesion, we modified the sentence generation prompt to exclude the prefix and conducted the sentence fusion step (step (c) in Fig. 2) using the ICL variant for both MDS and LFQA tasks, maintaining identical

---

[8]See examples of fine-grained attributions generated by our approach in Appendix E.

[9]See partially-attributed examples in Appendix H.

[10]This trend is also observed in our automatic evaluation, where extending sub-sentence attributing spans to their corresponding full sentences improves AUTOAIS. See Appendix D.

sentence-level plans. Starting from the second sentence to isolate the impact of the prefix, we examined 20 instances per task. Our analysis of 113 sentences reveals that approximately 30% were more cohesive when including the prefix. Specifically, 23% showed improved referential linkage, mostly attributed to better use of co-reference, and 7% demonstrated better discourse connector usage.[11] Additionally, our findings suggest that prefixes play an increasing role in cohesion as the output progresses (see Fig. 3), implying that cohesion gains importance as the text develops.[12]

6 Related Work

A few recent works have proposed several methods to generate text accompanied by attribution, with varying levels of granularity in the source and the generated text. In Long-Form Question Answering (LFQA), the LaMDA system (Thoppilan et al., 2022) provides attribution for the *entire* response in the form of URLs pointing to *entire* documents. In GopherCite (Menick et al., 2022), the *entire* response is attributed to fine-grained snippets from the documents. ALCE (Gao et al., 2023b) provides attribution for each generated sentence to one or multiple *entire* input documents. SEMQA (Schuster et al., 2023) is a recent multi-source semi-extractive dataset for LFQA, which encourages models to integrate factual spans copied verbatim from the input, along with free-text connectors to ensure coherence. This method is by design fine-grained both at the source and the response level using string matching. Yet, this semi-extractive paradigm is limited in scenarios requiring a higher level of abstraction, such as Multi-document Summarization (MDS). In contrast to all those approaches, our localized attribution achieves granularity both in the input and the output, without restricting the model to extractive text generation.

In information seeking, recent works have proposed to find attribution in post-hoc, by retrieving documents and matching each part of the answer to specific pieces of text from the retrieved documents (Bohnet et al., 2023; Gao et al., 2023a). These methods face a combinatorial challenge because the relevant snippets of text in the source documents need to be determined.

The field of attributed text generation is part of a broader endeavor to address hallucinations in models' outputs and increase their overall trustworthiness. In this context, much work has been done on characterizing the sources of hallucinations (Dziri et al., 2022; Rawte et al., 2023; Yao et al., 2023; Li et al., 2024) as well as on strategies to identify and mitigate them (Azaria and Mitchell, 2023; Manakul et al., 2023; Min et al., 2023; Slobodkin et al., 2023b; Mishra et al., 2024b), thereby advancing the reliability of model-generated content.

Another related line of research focuses on text-planning for improving output quality and reducing hallucinations (Gehrmann et al., 2018; Narayan et al., 2021, 2023; Zhao et al., 2020), by introducing an additional step of planning before the final text generation. This area of study includes efforts to decompose the generation process into distinct phases, with some works focusing on the content-selection step (Gehrmann et al., 2018; Cho et al., 2019; Ernst et al., 2021), while others focus on the content consolidation phase (Lebanoff et al., 2019, 2020; Slobodkin et al., 2022, 2023a; Hirsch et al., 2023). Our work also leverages this planning paradigm to improve the factuality of the output. However, unlike previous methods, our pipeline is carefully designed to also provide high-quality fine-grained attributions. Perhaps most related to our method is the work by Ernst et al. (2022) that divides the MDS process into detecting salient propositions, clustering them, and transforming them into distinct sentences. However, their method results in outputs that resemble bullet-point summaries, consisting of single-proposition sentences, rather than producing more fluid, multi-proposition sentences typical of natural text. In contrast, our method maintains the coherence and natural flow of generated texts without being restricted to single-proposition sentences.

Text-planning is also related to the Chain-of-Thought (CoT) approach that demonstrated improved performance on math problems (Wei et al., 2022), coding and programming puzzles (Chen et al., 2021; Schuster et al., 2021), and other reasoning tasks (Saparov and He, 2023). Both approaches build on the idea of decomposing the generation task to allow an initial reasoning and planning stage, before generating the predicted output. In CoT, the planning and output prediction are typically performed in a single generation call, whereas in text-planning the planning stage is separated and could be performed either by the same LM or using different models and operations. Therefore, CoT

---

[11] See Appendix I for examples.

[12] The observed plateau in cohesion levels in LFQA is attributed to its outputs generally not exceeding three sentences.

is more suitable for in-context learning whereas text-planning works well with finetuned specialized models. We explore both options in this work.

7 Conclusion

In this work, we present *"Attribute First, then Generate"*, a novel scheme for attributable grounded text generation, comprised of separating the generation process into multiple steps, such that attribution is achieved as a by-product of the generation. To the best of our knowledge, we are the first to provide such fine-grained attributions in both the generated text and the references. Through automatic and manual evaluations, we showed that our decomposed generation pipeline either matches or outperforms existing baselines in terms of task-specific metrics. Importantly, our approach yields orders of magnitude shorter attributions, shortening the fact-checking time by almost 50%, with little impact on the attribution accuracy. While this work focused on a particular implementation of content selection and generation, we encourage future explorations to extend our *"Attribute First, then Generate"* paradigm to other configurations. We hope that our research will motivate future work in this direction, improving the quality of text generation while also providing helpful concise attributions to external sources, thereby increasing models' trustworthiness and safe downstream use.

Limitations

Our Attribute First, Then Generate approach tends to use more computing resources and is also slower compared to direct end-to-end generation, as detailed in Appendix J. This is because it requires several steps, each involving a separate call to the model.

Furthermore, our pipeline is comprised of several interlinked components, which can potentially lead to error propagation. However, this potential issue is partially alleviated by the inherent robustness of the individual components. For instance, should the content selection mechanism identify a broad array of text fragments, including some that are irrelevant, subsequent stages in the process are designed to filter out such non-essential information.

Ethics Statement

Crowdsourcing. For our human evaluation we used the Amazon Mechanical Turk[13] (MTurk) crowdsourcing platform. Recruitment of annotators was conducted via email invitations dispatched from the MTurk platform, aimed at individuals who had previously demonstrated high performance in similar research endeavors. Compensation was provided not only for the actual annotation tasks but also for an initial onboarding phase, which included familiarization with the annotation guidelines and going over feedback from the authors. The pricing for each annotation instance was calibrated based on an estimated completion time, aiming to maintain a compensation rate of approximately 12$ per hour. Additionally, we closely monitored the actual time required for task completion and offered post-hoc reimbursements in compensation for instances where the required time exceeded our initial estimate by more than 5%.

Datasets licenses. The DUC and TAC datasets were acquired according to the required NIST guidelines (`duc.nist.gov`), and we do not publish any data from these datasets. The annotation data from Liu et al. (2023) is published with the MIT license. The Multi-News dataset (Fabbri et al., 2019) is published with a non-commercial license intended for research and educational purposes. Accordingly, our human evaluation annotations will also be limited to non-commercial use.

AI assistants AI assistants were exclusively employed to enhance the grammatical structuring of the text.

Acknowledgements

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

---

[13]`https://worker.mturk.com/`

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

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.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052.

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?

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

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.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. Summeval: Re-evaluating summarization evaluation.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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.

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

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Eran Hirsch, Valentina Pyatkin, Ruben Wolhandler, Avi Caciularu, Asi Shefer, and Ido Dagan. 2023. Revisiting sentence union generation as a testbed for text consolidation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7038–7058, Toronto, Canada. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. Learning to fuse sentences with transformers for summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4136–4142, Online. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne 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.

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

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

Aviya Maimon and Reut Tsarfaty. 2023. A novel computational and modeling foundation for automatic coherence assessment.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

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.

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

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024a. Fine-grained hallucination detection and editing for language models. *ArXiv*, abs/2401.06855.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024b. Fine-grained hallucination detection and editing for language models.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2022. Measuring attribution in natural language generation models.

Vipula Rawte, Prachi Priya, S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Amit Sheth, and Amitava Das. 2023. Exploring the relationship between llm hallucinations and prompt linguistic nuances: Readability, formality, and concreteness.

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.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Tal Schuster, Ashwin Kalyan, Alex Polozov, and Adam Kalai. 2021. Programming puzzles. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Tal Schuster, Adam D. Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William W. Cohen, and Donald Metzler. 2023. Semqa: Semi-extractive multi-source question answering.

Aviv Slobodkin, Avi Caciularu, Eran Hirsch, and Ido Dagan. 2023a. Don't add, don't miss: Effective content preserving generation from pre-selected text spans. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12784–12800, Singapore. Association for Computational Linguistics.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023b. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. 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.

Aviv Slobodkin, Ori Shapira, Ran Levy, and Ido Dagan. 2024. Multi-review fusion-in-context.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

A Datasets Preprocessing

Datasets statistics are reported in Table A.1. For LFQA, we utilize published human annotations of LLM responses with in-line citations (Liu et al., 2023). We use the existing train, dev, and test splits, and we apply a few filters on this dataset as follows. Responses from LLMs can contain statements that are not supported by the retrieved documents and

Method	Train	Eval	Test
EVALUATING (LIU ET AL., 2023)	393	44	45
MULTI-NEWS (FABBRI ET AL., 2019)	-	28	65
DUC & TAC (duc.nist.gov)	475	119	-

Table A.1: Our different datasets sizes used for training, development, testing and hyperparameter-tuning.

	Kendall-Tau		Spearman	
	τ	CI	τ	CI
TRUE	**0.4659**	**0.46-0.47**	**0.5215**	**0.51-0.53**
TRUE$_{prob.}$	0.3480	0.34-0.36	0.4182	0.41-0.43
TrueTeacher	0.4550	0.45-0.46	0.5130	0.50-0.52
TrueTeacher$_{prob.}$	0.3408	0.33-0.35	0.4152	0.41-0.42

Table A.2: Average Kendall-Tau and Spearman's rank correlations (τ) and their 95% confidence intervals (CI) for tested evaluation metrics against human judgment. "prob." refers to the variants averaging the respective model's probabilities of the label '1'.

are annotated with "Partial Support" or "No Support". We classify each statement as supported by the input documents if either one of its citations is annotated with "Full Support" or if the union of all citations is annotated with "Full Support". We remove examples from the dataset containing unsupported statements. In addition, the dataset also includes human annotations for perceived utility (on a Likert scale between 1 to 5). We only keep examples with perceived utility 4 and 5 ("Strongly Agree" and "Agree"). Lastly, the dataset contains evidence excerpts for each generated statement, which we use as citations. We remove examples from the dataset if we cannot align the evidence excerpt with the source documents.

For the training processes in the MDS setting, we use the alignment dataset from Ernst et al. (2022), automatically extracted from the DUC and TAC summarization datasets.[14] For the development and test splits, we use the alignment dataset from Ernst et al. (2024), extracted from the Multi-News summarization dataset (Fabbri et al., 2019). The dataset consists of 100 summaries with a corresponding set of source news articles, comprising an average of 3 articles per summary. Additionally, for each such pair of summaries and set of documents, the authors trained qualified annotators to identify aligning spans, following the approach delineated in Slobodkin et al. (2022).

B   Meta-Evaluation for AUTOAIS

Prior work that suggested using NLI models for automatic attribution evaluation (Bohnet et al., 2023; Gao et al., 2023b,a) has specifically advocated using the TRUE model (Honovich et al., 2022). Subsequent developments have introduced an enhanced NLI model, named TrueTeacher (Gekhman et al., 2023). In our study, to determine the most appropriate NLI model for the automatic assessment of attribution within the AUTOAIS framework, we conduct a correlation analysis between the outcomes of these models and our human-generated annotations. We also evaluate an alternative method that averages the models' probabilities for the entailing label ('1') across all sentences, rather than binary outcomes. To calculate correlation, we use both Kendall-Tau and Spearman's rank correlations, as suggested in (Deutsch et al., 2022). We also apply bootstrapping (Efron, 1987) by performing 1000 samplings of 150 instances (with repetition) and calculating correlation scores for each such subset. We report the average correlation and 95% confidence intervals for each metric. Our findings, presented in Table A.2, indicate that despite the advancements offered by the TrueTeacher model, the TRUE model still aligns more closely with human judgment. Furthermore, the approach of averaging probabilities for the label '1' correlates significantly worse for both models. Consequently, we opt to continue utilizing the TRUE model for our AUTOAIS evaluations.

C   Implementation Details

In this section, we describe the implementation details for all the component described in Section 3, both the ICL and fine-tuning approaches. ROUGE$_L$ and BERTSCORE were calculated with the evaluate library [15] (Lin, 2004).

For our ICL experiments, we use the Gemini-Pro model (Gemini Team Google, 2023), last updated on December 2023, which is available via API calls.[16] In all LFQA experiments, the query is included as part of the input for each component. For all subtasks and baselines, we perform a grid search on the optimal number of in-context examples (1-3 in the MDS setting and 1-4 in the LFQA setting) and temperature hyperparameter (0.1,0.3,0.5,0.7,0.9) by evaluating each combination on the respective development set. For all

---

[14]https://duc.nist.gov/

[15]https://huggingface.co/docs/evaluate/en/index
[16]https://ai.google.dev/models/gemini

fine-tuned variants, we fine-tune PRIMERA for both tasks. We also tested training LONGT5, but found PRIMERA performs significantly better. The number of parameters in the PRIMERA model is 447M. All fine-tuned models were trained for 10 epochs and the best model checkpoint was chosen by ROUGE$_L$, unless mentioned otherwise. The training of each model took approximately 1 hour on one Nvidia A100 80GB GPU. Hyperparameter training included 3 epochs for each model and used the following sweep configuration: learning rate (max: 5e-4; min: 5e-8), warmup steps (max: 300; min: 0), weight decay (0.0, 0.2, 0.5).

C.1 Content Selection

ICL In the ICL method, we present the model with the source texts and guide it to copy specific segments from those texts (see Figs. B.1 and B.5 for prompt examples in the MDS and LFQA settings, respectively). We use a designated *"<SPAN_DELIM>"* markup as a delimiter between such consecutive spans, for better parsing. Within the MDS framework, the model is tasked with identifying salient excerpts, whereas in the context of LFQA, the model is also provided with a query, and is asked to find the parts of the texts that are relevant to answering this question. Further, following our analyses of each setting's development sets, we restrict the model to an average and maximum number of words to choose (an average of 200 and 100 words and a maximum of 900 and 200 words for MDS and LFQA, respectively). Lastly, to manage source texts that exceed the model's input size limit, we truncate the fewest necessary sentences from those texts to ensure they fit within the specified limit.[17] After performing hyperparameter tuning, in the MDS we set $T = 0.1$ and the number of in-context examples to 2, while in LFQA we set $T = 0.3$ and the number of examples to 4.

Fine-tuned Examples are provided in Figs. B.9 and B.12. The input to the model is the set of all documents, concatenated with a special markup "<doc-sep>", which was also used in pre-training in PRIMERA. To fit the input in a context length of 4096, we trim each document to the length of 4096 divided by the number of documents in each example. The output is a set of highlights separated with a new special markup "<highlights-separator>". We run hyperparameter-tuning on

---

[17]We use spaCy (Honnibal and Montani, 2017) to separate documents into sentences.

each setting based on the development set. The hyperparameters used for MDS are warmup steps of 229, weight decay of 0.5, and a learning rate of 4e-4. Conversely, the hyperparameters used for LFQA are warmup steps of 234, weight decay of 0, and a learning rate of 1e-4. We chose the best model checkpoint by calculating Intersection-over-Union (IoU) between the gold highlights spans and the predicted highlights spans.

C.2 Generation Planning

ICL In the ICL setting, we present the model with the source text where we insert the mark-ups *"<highlight_start>"* and *"<highlight_end>"* before and after each pre-selected span, (see Figs. B.2 and B.6 for the MDS and LFQA settings, respectively). We also enumerate the highlights across all documents, and for the LFQA setting, we also add the query. The model is directed to cluster highlighted segments that are likely to co-occur in a sentence in the same order it would then merge them into an output. We discovered that providing the model with task-specific motivation enhances its performance; for instance, in MDS, the objective is to fuse highlights into a summary, whereas in LFQA, it aims to formulate a response to the query. To facilitate the parsing of the model's response, we encourage it to generate its response as a list of single-value dictionaries, with "cluster" as the key and a list of indices corresponding to the highlights within the current cluster as the value. For hyperparameter tuning, we divide the gold "highlights" into pairs. Pairs that correspond to the same target output sentence are categorized as positive, whereas others are considered negative. Subsequently, we evaluate the F1 score to assess the alignment between the predicted and gold clusters. This involves calculating the True Positives, False Negatives, and False Positives for the highlights pairs, based on the predicted clusters. Following our analyses, we opt for 2 examples and $T = 0.5$ in the MDS setting, and 1 example and $T = 0.1$ in the LFQA setting.

Fine-tuned Examples are provided in Figs. B.10 and B.13. There are two variants for generation planning: generative clustering and joint. In the generative clustering approach, the input to the model is the set of all highlights, concatenated with a special markup "<highlight-separator>", which is the same as the output from the content selection model. The output is then a set of clusters,

where each cluster is separated by a special markup "<cluster-separator>", and inside each cluster we have a set of highlights in the same format as the input. We run hyperparameter-tuning on each setting based on the development set. The hyperparameters used for MDS are warmup steps 0, weight decay 0, and learning rate 5e-5. The hyperparameters used for LFQA are warmup steps 298, weight decay 0, and learning rate 2e-4. At inference time, we used a beam size of 1. We use the same logits processor described in Appendix C.1 to enforce the model to generate valid highlights.

In the joint approach, the input to the model is the original source documents, similar to the input in the content selection model. In contrast, the output is the clustered highlights. Similar to the content selection model, we use the same logits processor to enforce generating valid highlights. The hyperparameters used for LFQA are warmup steps 298, weight decay 0, and learning rate 2e-4.

C.3 Attributed Fused Output

ICL For the ICL variant, we iteratively show the model only the current cluster's highlights, in the same manner, we do so in the generation planning ICL experiments (see Appendix C.2). Additionally, we present the model with what it generated to that point and guide it to generate the next sentence in the paragraph, which should cover all and only the highlighted spans (see Figs. B.3 and B.7). We use 3 examples and $T = 0.3$ for the MDS setting, and 2 examples with $T = 0.1$ for the LFQA setting.

For the Chain-of-Thought approach, which merges generation planning with the fusion steps, we present the model with the "highlighted" source texts and ask it to fuse those highlights into a coherent output. Additionally, we guide the model, both in the instructions and through in-context demonstrations to iteratively cluster highlights and fuse them into a sentence, before generating the final output (see Figs. B.4 and B.8). We find that 3 in-context examples and $T = 0.1$ work best for MDS, while for LFQA it is 1 example and $T = 0.3$ that yield the best results on the development set.

Fine-tuned Examples are provided in Figs. B.11 and B.14. The input to the attributable fusion model is a single cluster of highlights, where highlights are separated by a special markup. The output is then a single sentence of the response. This model is called iteratively, and at each iteration, we include the prefix of the already generated re-

sponse. The output also includes citations for every sentence in the generated response based on the highlights used to generate the particular sentence. The hyperparameters for the fusion model are for MDS are warmup steps 54, weight decay 0.2, and learning rate 1e-4. The hyperparameters used for LFQA are warmup steps 175, weight decay 0.5, and learning rate 2e-5.

C.4 End-to-end Generation

We include as a baseline end-to-end generation without decomposing the process into several steps.

ICL For the ICL setting, we instruct the model to perform the corresponding setting's task, namely summarization for MDS, and answering a query based on the supplied document for LFQA. In addition to the varying number of in-context examples, we also evaluate the model in a zero-shot setting. Our analyses show that for MDS, 1 example and $T = 0.5$ give the best performance, while for LFQA 3 examples and $T = 0.7$ are optimal.

Fine-tuned The input to the model is the set of all documents, concatenated with a special markup "<doc-sep>", which was also used in pre-training in PRIMERA. Similar to our content selection, we trim each document to the length of 4096 divided by the number of documents in each example. The output is simply the full response, a summary for MDS, or an answer for LFQA. The hyperparameters used for both MDS and LFQA are warmup steps 27, weight decay 0.2, and learning rate 4e-5.

C.5 ALCE Variant

As another baseline, we also implement the ALCE approach (Gao et al., 2023b) for attributed generation, where the model is guided to jointly generate a text with in-line citations, pointing to the supporting source documents. We use the same prompt as in the original paper. Additionally, we find that for both settings, 2 examples and $T = 0.5$ yield the best results on the respective development sets.

C.6 Constrained Decoding

Both the content selection component and the planning component are designed to exactly copy text spans from source documents. To guide the generation to precisely follow the lexical structure of the input's text, we employ a constrained decoding strategy for the fine-tuned models. This strategy

	Method	AUTOAIS↑	AUTOAIS$_{full\ sentence}$↑
ICL	ALCE	88.7	88.7
	ATTR. FIRST	79.5	80.6
	ATTR. FIRST$_{CoT}$	72.8	75.0
FT	ATTR. FIRST	64.4	80.1
	ATTR. FIRST$_{joint}$	48.9	83.9

Table C.1: MDS results comparing our reported AutoAIS to a variant of AutoAIS where the full-sentences containing the highlights are used as premise. The difference between the two AutoAIS variants is that the former might contain non-grammatical sentences.

	Method	AUTOAIS↑	AUTOAIS$_{full\ sentence}$↑
ICL	ALCE	49.8	49.8
	ATTR. FIRST	78.7	80.9
	ATTR. FIRST$_{CoT}$	89.3	91.5
FT	ATTR. FIRST	52.7	56.0
	ATTR. FIRST$_{joint}$	44.9	69.2

Table C.2: LFQA results comparing our reported AutoAIS to a variant of AutoAIS where the full-sentences containing the highlights are used as premise. The difference between the two AutoAIS variants is that the former might contain non-grammatical sentences.

relies on a logits processor[18], allowing the model to avoid certain tokens by assigning their scores to minus infinity. Specifically, the model is restricted from generating n-grams absent from the source documents. Additionally, to optimize the models for producing concise and informative highlights, we impose additional constraints, including setting minimum thresholds on the number of words in each highlight and the overall number of highlights generated. Based on hyperparameter-tuning we set a minimum of three words for each highlight in both MDS and LFQA, and a minimum number of 30 highlights for MDS and 5 highlights for LFQA. These constraints contribute significantly to optimizing the performance and reliability of the fine-tuned generation models. Lastly, in the planning component, the model organizes and forms clusters, each encapsulating highlights from the given input. To prevent excessively large clusters, we impose a maximum threshold of two highlights per cluster. This facilitates the generation of coherent sentences from each cluster.

D Full-sentence AUTOAIS

Tables C.1 and C.2 present the results on the full-sentence AutoAIS vs. regular AutoAIS

E Fine-grained Attribution Examples

Table D.1 shows fine-grained attributions examples, generated by ATTR. FIRST$_{CoT}$ in the MDS setting.

F Human Study

For our human study, we followed the Controlled Crowdsourcing Methodology (Roit et al., 2020) to detect qualified annotators via Amazon Mechanical Turk.[19] We reached out to potential candidates, who had exhibited high proficiency in similar tasks in the past, through email invitations sent from the MTurk platform. During onboarding, candidates reviewed the annotation guidelines and completed a closed qualification round consisting of five annotation tasks. This process, overseen by an author of the paper who provided individualized feedback, required approximately 20 minutes to complete, with participants compensated 4.5$ for their efforts. Ultimately, six out of eleven candidates were selected for the actual evaluations, with compensation rates set at 2$ per ALCE instance and 1$ otherwise, reflecting the varying complexity of the instances. To ensure fair compensation, we also monitored the actual time required to complete each instance and offered post-hoc reimbursements when this time exceeded our initial estimates by more than 5%.

Moreover, to evaluate the consistency of attribution annotations, we computed the inter-annotator agreement (IAA) using Fleiss' Kappa coefficient (Fleiss, 1971), based on a sample of 101 sentences and their attributions from 23 instances across both models, each annotated by 3 workers.[20] The resulting Kappa coefficient was 0.37, indicating a fair level of agreement (Landis and Koch, 1977).

G Annotation Interface

Screenshots from our annotation interface are presented in Figs. E.1 and E.2. Our interface mirrors the design from Liu et al. (2023), adapted to support citations with highlights. In addition, the documents are presented within the interface rather than via external links. Within this interface, annotators assess each sentence to determine the level of support provided by its accompanying citations, choosing one of the following options: "Full Support", "Partial Support", "No Support", "Citation Has Support but also Refutes Statement", "Statement is Unclear, Can't Make Judgement". Finally, when calculating the final AIS score, we

---

[18]https://huggingface.co/docs/transformers/internal/generation_utils#logitsprocessor

[19]www.mturk.com

[20]We treat annotators' scores as binary, akin to Appendix G.

	Prediction
1	**Generated sentence:** Hawaii's tourism industry is worth $14 billion, and hotels and resorts are offering special packages to attract same-sex couples. [1] **[1]** . . . There's a large number of firms that are specialized in the marriage business." In Hawaii, tourism is a $14 billion industry. Now, the island chain is positioning itself for a spike in visitors - among them, Honolulu-based hotel chain Aqua Hospitality. It already offers LGBT travel deals, including the "out and proud" package and the "civil unions romance special." Bigger beachfront resorts are also getting in on the action. The Sheraton Waikiki is the only hotel on Oahu providing on-site marriage licenses. General Manager Kelly Sanders says now the hotel is running a new ad campaign in mainland LGBT publications . . .
2	**Generated sentence:** Patten provided lobbying and consulting services to a Ukrainian oligarch and his political party, but failed to register as a foreign agent as required by law. [1,2] **[1]** . . . From 2014, Patten provided a "prominent" Ukrainian oligarch who isn't named in court papers and his Opposition Bloc political party with lobbying and consulting services, according to the criminal information. A company Patten co-owned with a Russian national received more than $1 million for the work, the U.S. said. Details about the Ukrainian oligarch match those of Sergei Lyovochkin, an Opposition Bloc leader whom prosecutors have previously identified as funding Manafort's work in Ukraine. He couldn't immediately been reached for comment. As part of his lobbying work, Patten violated the Foreign Agents Registration Act by not disclosing the work to the U.S., prosecutors said. No date has been set for his sentencing . . . **[2]** . . . Prosecutors say Patten, who formed a consulting company with a person identified only as "Foreigner A," worked to set up meetings with members of Congress and also drafted talking points for Capitol Hill meetings. The goal, prosecutors say, was to influence U.S. policy, but they say Patten never filed under the Foreign Agents Registration Act. The law is aimed at promoting transparency about lobbying efforts in the United States. . . .
3	**Generated sentence:** On Tuesday, a shooting at a Home Depot store in North Dallas left three people injured: two police officers, Crystal Almeida and Rogelio Santander, and a Home Depot loss-prevention officer, Scott Painter. [1,2,3] **[1]** . . . The wounded officer is Crystal Almeida, 26. The third victim was identified as Scott Painter, a Home Depot loss-prevention officer. Almeida and Painter were still in critical condition on Wednesday, but they were making "remarkable recoveries," Dallas Police Chief U. Renee Hall said at a news conference Wednesday morning. . . . **[2]** . . . Along with Officer Santander, Officer Crystal Almeida and the Home Depot security guard, now identified as Scott Painter, all underwent surgery after the shooting. Officer Almeida and Mr. Painter both remain hospitalized. During the press conference Chief Hall said, "We are happy to report that Officer Crystal Almeida and our loss-prevention officer Scott Painter is making remarkable recovery. They are still in critical condition but we are optimistic about what we're seeing with them right now." The man who police say shot the officers and the security guard - Armando Luis Juarez - woke up behind bars today. After the shooting, a manhunt, and late night police chase, Juarez was taken into custody and ultimately transferred to the Dallas County Jail very early Wednesday morning. Officers Santander and Almeida both joined the Dallas Police Department there years ago, assigned to the Northeast Division. . . . **[3]** . . . Mayor Mike Rawlings was presiding over a city council meeting when he announced the death of Rogelio Santander, a member of the police force for three years. Santander, officer Crystal Almeida and a loss-prevention officer for Home Depot were shot Tuesday by a man identified by police as 29-year-old Armando Luis Juarez. The two officers and the store loss-prevention officer underwent surgery for their injuries after the shooting in the north of the city, Dallas Police Chief U. Renee Hall said late Tuesday. Almeida and the loss-prevention officer, who hasn't been identified, were in critical condition Wednesday. . . .

Table D.1: Attribution examples for several sentences generated by the ATTR. FIRST$_{\text{CoT}}$ variant for the MDS settings.

treat instances labeled as "Partial Support" as *non-attributed*, while instances classified as "Citation Has Support but also Refutes Statement" are labeled as *attributed*, akin to Gao et al. (2023a).

H Partially-attributing Examples

Table H.1 shows a few attribution examples labeled by human annotators as partially-supporting.

I With Prefix Versus without Prefix

Table H.2 presents several instances from our test set when the iterative sentence-level fusion was only conditioned on each iteration's cluster, versus when it was also conditioned on the prefix.

J Latency and Total Number of Input and Output Tokens

In Table H.3 we present the average number of tokens processed in the input and also the average number of generated tokens for each of our pipeline's subtasks, and also for our baselines. Additionally, in Table H.4 we present the average time in seconds it takes to run each of our pipeline's subtasks, as well as our baselines.

K Post-generation Attribution

We also investigated a post-generation attribution variant, where first an unattributed output was generated, and then relevant attribution was retrieved. To that end, we first generated the task-specific output, using our baseline from Section 4.2. Then, to obtain attributions, we adapted the RARR algorithm, as outlined by Gao et al. (2023a), to suit our context. In contrast to its original application of retrieving relevant documents, our adaptation employs the grounding texts directly. The process involves generating queries from each generated sentence about the various facets it encompasses, identifying 4-sentence paragraphs within the source texts that correspond to each query through a query-document relevance model, and answering query based on these paragraphs and the output sentence. The comparisons of these responses determine the selection of attributing paragraphs; only those that either entail or contradict are retained as valid attributions. The results, as detailed in Table K.1, show that this method produced both longer and less accurate attributions compared to those derived from our approach.

	Prediction
1	**Output:** Despite economic challenges and **low approval ratings**, Obama's campaign finance report shows he can still attract major cash. **Attribution:** … <mark>The president's campaign finance report shows he can still pull in major cash despite a stagnant economy</mark>, dipping approval ratings and grumblings among some liberal supporters that he has not done enough for their cause. …
2	**Output: The Sheraton Waikiki is the only hotel on Oahu that provides on-site marriage licenses**, and it is running a new ad campaign in mainland LGBT publications. **Attribution:** …The Sheraton Waikiki is the only hotel on Oahu providing on-site marriage licenses. <mark>General Manager Kelly Sanders says now the hotel is running a new ad campaign in mainland LGBT publications.</mark> …
3	**Output:** Tarek and his wife Christina tried for two years to have a son, going through two failed IVF attempts and **a miscarriage** before welcoming their son Brayden James in August. **Attribution:** … <mark>The couple, who are already parents to daughter Taylor, tried for more than two years to get pregnant with their son, going through two failed attempts at IVF</mark> - including one resulting in a miscarriage at eight weeks, just before Tarek's diagnosis. <mark>Now in remission, the new dad, who welcomed son Brayden James with Christina, 32, in August, is positive about his family's future."</mark> …

Table H.1: Example predictions of the ATTR. FIRST$_{CoT}$ which were annotated with partial support. The information inferred from the words in bold is not supported by the highlights provided by the model, but is supported by sentences in close proximity.

	Prediction
1	**w/ prefix:** The Ohio River starts at the confluence of the Allegheny and Monongahela Rivers in Pittsburgh, Pennsylvania, and flows 981 miles to its ending point in Cairo, Illinois. **The river** flows in a southwesterly direction from western Pennsylvania to its mouth on the Mississippi River at the southern tip of Illinois. **w/o prefix:** The Ohio River starts at the confluence of the Allegheny and Monongahela Rivers in Pittsburgh, Pennsylvania, and flows 981 miles to its ending point in Cairo, Illinois. **The Ohio River** is a 981-mile (1,579 km) long river in the United States.
2	**w/ prefix:** Idina Menzel and Taye Diggs, who starred together in the Broadway musical "Rent," have decided to end their marriage. **Their** main concern is their son. **The couple** , who met while starring in the Broadway musical "Rent," have a 4-year-old son, Walker Nathaniel. **Diggs and Menzel** , both 42, met as costars in the original production of the Broadway musical Rent in the mid-1990s. They worked at their marriage, going through tough times like everyone else, but they loved each other very much. **w/o prefix:** Idina Menzel and Taye Diggs, who have been married for 10 years, have a 4-year-old son named Walker. Idina Menzel and Taye Diggs have asked for privacy during this time, as their main concern is for their son. **Taye Diggs and Idina Menzel** have decided to separate after ten years of marriage, but their main concern is for their son. **Taye Diggs and Idina Menzel** , who met in the 1990s while starring in the Broadway musical Rent, have asked for privacy during this difficult time. **Taye Diggs and Idina Menzel** , who met in the mid-1990s while starring in the original production of the Broadway musical Rent, married in 2003 and had a son, Walker, in September 2009.
3	**w/ prefix:** By exposing students to different religious texts, practices, cultures, and social teachings, they can gain a deeper understanding of the world around them and develop critical thinking skills. **However ,** some Americans believe that teaching religious doctrine in public schools infringes on the First Amendment right to free exercise of religion. **w/o prefix:** By exposing students to different religious texts, practices, cultures, and social teachings, they can gain a deeper understanding of the world around them and develop critical thinking skills. Some Americans believe that excluding religious sentiment from public schools infringes on the First Amendment right to free exercise of religion.

Table H.2: Example predictions of the ICL ATTR. FIRST model when prompted with prefix and without prefix.

		Content Selection		Generation Planning		Sentence Fusion		CoT		End-to-end		ALCE	
		Input	Output	Input	Output	Input	Output	Input	Output	Input	Output	Input	Output
ICL	MDS	6879 (± 1893)	551 (± 467)	6425 (± 1728)	11 (± 22)	17003 (± 13203)	133 (± 80)	10273 (± 1558)	478 (± 220)	6879 (± 1893)	177 (± 55)	7341 (± 1974)	193 (± 44)
ICL	LFQA	10727 (± 3206)	142 (± 161)	2536 (± 1651)	9 (± 13)	3734 (± 1784)	45 (± 28)	2962 (± 635)	127 (± 89)	10727 (± 3206)	91 (± 85)	8476 (± 3809)	95 (± 89)
FT	MDS	1661 (± 116)	456 (± 30)	1041 (± 75)	456 (±30)	348 (± 7)	29 (± 0.6)	-	-	1661 (± 116)	283 (± 9)	-	-
FT	LFQA	2883 (±162)	121 (±11)	348 (±36)	121 (±11)	209 (±18)	27 (+ -1)	-	-	2883 (± 162)	60 (± 5)	-	-

Table H.3: Average number of tokens (and standard deviation) for each of the steps in our pipeline, as well as our two baselines.

		Content Selection	Generation Planning	Sentence Fusion	CoT	End-to-end	ALCE
ICL	MDS	33.1	5.2	5.6	17.1	7.0	9.2
ICL	LFQA	6.3	4.7	6.8	3.3	6.1	6.7
FT	MDS	68.8	5.7	2.4	-	1.7	-
FT	LFQA	10.6	3.4	1.6	-	1.2	-

Table H.4: Average time (in seconds) each of the steps in our pipeline takes per instance, as well as for our two baselines.

Task	ROUGE$_L$	BERTSCORE	AUTOAIS	LENGTH	NO ATT. (%)
MDS	19.0	86.1	67.8	137.7	0.7
LFQA	33.1	89.6	50.2	149.0	9.6

Table K.1: Results of the post-generation attribution approach, adapted from the RARR algorithm, on the Multi-document summarization (MDS) and the Long-form Question-answering (LFQA) tasks.

```
 1  In this task, you are presented with several documents, which need to be summarized. As an
        intermediate step, you need to identify salient content within the documents. For each
        document, copy verbatim the salient spans, and use <SPAN_DELIM> as a delimiter between
        each consecutive span. IMPORTANT: The output must be of the format Document [<DOC_ID>]:
        <SPAN_DELIM>-delimited consecutive salient spans. IMPORTANT: Each salient content must
        be a single consecutive verbatim span from the corresponding passages. IMPORTANT: make
        sure the total number of copied words (from all documents) is around 200 words, and not
        more than 900.
 2
 3  Document [1]: The United States on Wednesday warned US citizens of a possible spike in
        "terrorist activity" in Colombia in the coming week as the left-wing Revolutionary
        Armed Forces of Colombia (FARC) mark the one-year anniversary of the death of one of
        their commanders ...
 4  Document [2]: Colombia's hardline government said Friday it is willing to meet with
        members of the country's main leftist rebel group in an unprecedented offer aimed at
        freeing dozens of rebel hostages, including three Americans. Before any face-to-face
        meeting can take place, however, the rebels must agree to swap 15 hostages for 15
        rebels jailed on minor charges, said Peace Commissioner Luis Carlos Restrepo ...
 5  ...
 6
 7  Answer:
 8  Document [1]: The United States on Wednesday warned US citizens <SPAN_DELIM> mark the
        one-year anniversary of the death of one of their commanders <SPAN_DELIM> ...
 9  Document [2]: Colombia's hardline government said <SPAN_DELIM> it is willing to meet with
        members of the country's main leftist rebel group <SPAN_DELIM> ...
10  ...
11
12  In this task, you are presented with several documents, which need to be summarized. As an
        intermediate step, you need to identify salient content within the documents. For each
        document, copy verbatim the salient spans, and use <SPAN_DELIM> as a delimiter between
        each consecutive span. IMPORTANT: The output must be of the format Document [<DOC_ID>]:
        <SPAN_DELIM>-delimited consecutive salient spans. IMPORTANT: Each salient content must
        be a single consecutive verbatim span from the corresponding passages. IMPORTANT: make
        sure the total number of copied words (from all documents) is around 200 words, and not
        more than 900.
13
14
15  Document [1]: Voters in 11 states will pick their governors tonight, and Republicans
        appear on track to increase their numbers by at least one, with the potential to
        extend their hold to more than two-thirds of the nation's top state offices ...
16  Document [2]: Eight of the gubernatorial seats up for grabs today are now held by
        Democrats; three are in Republican hands. Republicans currently hold 29
        governorships, Democrats have 20 ...
17  ...
18
19  Answer:
```

Figure B.1: Example prompt for the content selection subtask in the MDS setting.

```
  1   In this task, you are presented with several passages, where some parts are "highlighted"
          (namely, there are <highlight_start> and <highlight_end> tokens before and after each
          such span). The goal is to fuse all those highlights into a single summary. As an
          intermediate step, you need to cluster highlights that can be merged into a sentence
          (namely, each cluster will be later merged into one sentence). Make sure the clusters
          are in the same order you would then write the corresponding summary sentences.
          IMORTANT: make sure there are at least 7 clusters, and no more than 5-6 highlights per
          cluster. IMPORTANT: The output must be of the format [{"cluster":[comma-delimited
          highlights indices]}]
  2
  3   Document [1]: <highlight_start>The United States on Wednesday warned US
          citizens<highlight_end> of a possible spike in "terrorist activity" in Colombia in
          the coming week as the left-wing Revolutionary Armed Forces of Colombia (FARC)
          <highlight_start>mark the one-year anniversary of the death of one of their
          commanders<highlight_end> ...
  4   Document [2]: <highlight_start>Colombia's hardline government said<highlight_end> Friday
          <highlight_start>it is willing to meet with members of the country's main leftist
          rebel group<highlight_end> in an unprecedented offer aimed at freeing dozens of rebel
          hostages, including three Americans. Before any face-to-face meeting can take place,
          however, the rebels must agree to swap 15 hostages for 15 rebels jailed on minor
          charges, said Peace Commissioner Luis Carlos Restrepo ...
  5   ...
  6
  7   The highlighted spans are:
  8   Document [1]:
  9    1. The United States on Wednesday warned US citizens
 10    2. mark the one-year anniversary of the death of one of their commanders
 11   ...
 12   Document [2]:
 13    7. Colombia's hardline government said
 14    8. it is willing to meet with members of the country's main leftist rebel group
 15   ...
 16
 17   Answer: The highlighted spans are clustered as follows:
 18   [{"cluster":[1,2,5]}, {"cluster":[7,8]} ...]
 19
 20   In this task, you are presented with several passages, where some parts are "highlighted"
          (namely, there are <highlight_start> and <highlight_end> tokens before and after each
          such span). The goal is to fuse all those highlights into a single summary. As an
          intermediate step, you need to cluster highlights that can be merged into a sentence
          (namely, each cluster will be later merged into one sentence). Make sure the clusters
          are in the same order you would then write the corresponding summary sentences.
          IMORTANT: make sure there are at least 7 clusters, and no more than 5-6 highlights per
          cluster. IMPORTANT: The output must be of the format [{"cluster":[comma-delimited
          highlights indices]}]
 21
 22
 23   Document [1]: <highlight_start>Voters in 11 states will pick their governors
          tonight<highlight_end>, and Republicans appear on track to increase their numbers by
          at least one, with the potential to extend their hold to more than two-thirds of the
          nation's top state offices ...
 24   Document [2]: Eight of the gubernatorial seats up for grabs today are now held by
          Democrats; three are in Republican hands. <highlight_start>Republicans currently hold
          29 governorships, Democrats have 20<highlight_end> ...
 25   ...
 26
 27   The highlighted spans are:
 28   Document [1]:
 29    1. Voters in 11 states will pick their governors tonight
 30   ...
 31   Document [2]:
 32    5. Republicans currently hold 29 governorships, Democrats have 20
 33   ...
 34   Answer: The highlighted spans are clustered as follows:
```

Figure B.2: Example prompt for the clustering subtask in the MDS setting.

```
 1  In this task, you are presented with several passages, where some parts are "highlighted"
       (namely, there are <highlight_start> and <highlight_end> tokens before and after each
       such span). You are also presented with a prefix of a paragraph. You job is to generate
       the next sentence in the paragraph, that covers all and only the "highlighted" spans.
       Make sure it connects well with the prefix, and that it covers all and only the
       "highlighted" spans.
 2
 3  Document [1]: <highlight_start>The United States on Wednesday warned US
       citizens<highlight_end> of a possible spike in "terrorist activity" in Colombia in
       the coming week as the left-wing Revolutionary Armed Forces of Colombia (FARC)
       <highlight_start>mark the one-year anniversary of the death of one of their
       commanders<highlight_end> ...
 4  Document [2]: Colombia's hardline government said Friday it is willing to meet with
       members of the country's main leftist rebel group in an unprecedented offer aimed at
       freeing dozens of rebel hostages, including three Americans ...
 5  ...
 6
 7  Prefix: Colombia's strict government agreed to meet with the Country's main leftist rebel
       group.
 8
 9  The highlighted spans are:
10  Document [1]:
11   1. The United States on Wednesday warned US citizens
12   2. mark the one-year anniversary of the death of a terrorist commanders, American
         citizens recieved a warning from their government
13  ...
14  Answer:
15  The next sentence is:
       Following the one-year anniversary of the death of one of the group's commanders,
       American citizens received a warning from their government on Wednesday.
16
17  In this task, you are presented with several passages, where some parts are "highlighted"
       (namely, there are <highlight_start> and <highlight_end> tokens before and after each
       such span). You are also presented with a prefix of a paragraph. You job is to generate
       the next sentence in the paragraph, that covers all and only the "highlighted" spans.
       Make sure it connects well with the prefix, and that it covers all and only the
       "highlighted" spans.
18
19  Document [1]: <highlight_start>Voters in 11 states will pick their governors
       tonight<highlight_end>, and Republicans appear on track to increase their numbers by
       at least one, with the potential to extend their hold to more than two-thirds of the
       nation's top state offices ...
20  Document [2]: Eight of the gubernatorial seats up for grabs today are now held by
       Democrats; three are in Republican hands. <highlight_start>Republicans currently hold
       29 governorships, Democrats have 20 ...
21  ...
22
23  Prefix: 29 governorships are currently held by Republicans, vs. Democrat's 20.
24
25  The highlighted spans are:
26  Document [1]:
27   1. Voters in 11 states will pick their governors tonight
28  ...
29  Answer:
30  The next sentence is:
```

Figure B.3: Example prompt for iterative, sentence-by-sentence generation approach in the MDS setting.

```
 1  In this task, you are presented with several passages, where some parts are "highlighted"
        (namely, there are <highlight_start> and <highlight_end> tokens before and after each
        such span). Your job is to generate a coherent summary that covers all and only the
        "highlighted" spans. Build the summary iteratively, where each time you cluster
        highlights to build the next sentence. Make sure ALL highlights are included.
        IMPORTANT: make sure that the final summary consists of approximately 200 words and at
        least 7 sentences. IMPORTANT: The clustering steps must be of the form: "Highlights
        {comma-delimited highlights indices} are combined to form sentence
        {sentence_id}:sentence_text". Additionally, before generating the final summary, say
        "So the final summary is:".
 2
 3  Document [1]: <highlight_start>The United States on Wednesday warned US
        citizens<highlight_end> of a possible spike in "terrorist activity" in Colombia in
        the coming week as the left-wing Revolutionary Armed Forces of Colombia (FARC)
        <highlight_start>mark the one-year anniversary of the death of one of their
        commanders<highlight_end> ...
 4  Document [2]: <highlight_start>Colombia's hardline government said<highlight_end> Friday
        <highlight_start>it is willing to meet with members of the country's main leftist
        rebel group<highlight_end> in an unprecedented offer aimed at freeing dozens of rebel
        hostages, including three Americans ...
 5  ...
 6  The highlighted spans are:
 7  Document [1]:
 8   1. The United States on Wednesday warned US citizens
 9   2. mark the one-year anniversary of the death of one of their commanders
10  ...
11  Document [2]:
12   7. Colombia's hardline government said
13   8. it is willing to meet with members of the country's main leftist rebel group
14  ...
15  Answer:
16  The highlighted spans are combined as follows:
17  Highlights 7,8 are combined to form sentence 1: Colombia's strict government agreed to
        meet with the Country's main leftist rebel group.
18  ...
19  So the final summary is:  Colombia's strict government agreed to meet with the Country's
        main leftist rebel group. ...
20
21  In this task, you are presented with several passages, where some parts are "highlighted"
        (namely, there are <highlight_start> and <highlight_end> tokens before and after each
        such span). Your job is to generate a coherent summary that covers all and only the
        "highlighted" spans. Build the summary iteratively, where each time you cluster
        highlights to build the next sentence. Make sure ALL highlights are included.
        IMPORTANT: make sure that the final summary consists of approximately 200 words and at
        least 7 sentences. IMPORTANT: The clustering steps must be of the form: "Highlights
        {comma-delimited highlights indices} are combined to form sentence
        {sentence_id}:sentence_text". Additionally, before generating the final summary, say
        "So the final summary is:".
22
23  Document [1]: <highlight_start>Voters in 11 states will pick their governors
        tonight<highlight_end>, and Republicans appear on track to increase their numbers by
        at least one, with the potential to extend their hold to more than two-thirds of the
        nation's top state offices ...
24  Document [2]: Eight of the gubernatorial seats up for grabs today are now held by
        Democrats; three are in Republican hands. <highlight_start>Republicans currently hold
        29 governorships, Democrats have 20<highlight_end> ...
25  ...
26  The highlighted spans are:
27  Document [1]:
28   1. Voters in 11 states will pick their governors tonight
29  ...
30  Document [2]:
31   5. Republicans currently hold 29 governorships, Democrats have 20
32  ...
33  Answer:
34  The highlighted spans are combined as follows:
```

Figure B.4: Example prompt for the CoT approach in the MDS setting.

3332

```
 1   In this task, you are presented with a question and several documents. The goal is to
         answer the question based on the given documents. As an intermediate step, you need to
         identify the minimal number of spans within the documents that answer the question. For
         each document, copy verbatim those spans, and use <SPAN_DELIM> as a delimiter between
         each consecutive span. IMPORTANT: The output must be of the format Document [<DOC_ID>]:
         <SPAN_DELIM>-delimited consecutive spans. IMPORTANT: Each such span must be a single
         consecutive verbatim span from the corresponding passages. IMPORTANT: make sure the
         total number of copied words (from all passages) is around 100 words, and not more than
         200.
 2
 3   Question: What was the color of the carpet at the Oscars this year, and why wasn't it red?
 4
 5   Document [1]: ... tasked to reimagine the carpet color, opting to break tradition so the
         talent could better transition from daytime arrivals to the evening, according to the
         Hollywood Reporter ...
 6   Document [2]: ... rolling out the red carpet. The celebrated carpet at the Oscars -- the
         place where stars show off their most elegant and outrageous fashions -- will be a
         "champagne" color this year. It will be the first time ...
 7   ...
 8
 9   Answer:
10   Document [1]: reimagine the carpet color <SPAN_DELIM> so the talent could better
         transition from daytime arrivals to the evening <SPAN_DELIM> ...
11   Document [2]: The celebrated carpet at the Oscars <SPAN_DELIM> will be a "champagne" color
         this year. <SPAN_DELIM> ...
12   ...
13
14   In this task, you are presented with a question and several documents. The goal is to
         answer the question based on the given documents. As an intermediate step, you need to
         identify the minimal number of spans within the documents that answer the question. For
         each document, copy verbatim those spans, and use <SPAN_DELIM> as a delimiter between
         each consecutive span. IMPORTANT: The output must be of the format Document [<DOC_ID>]:
         <SPAN_DELIM>-delimited consecutive spans. IMPORTANT: Each such span must be a single
         consecutive verbatim span from the corresponding passages. IMPORTANT: make sure the
         total number of copied words (from all passages) is around 100 words, and not more than
         200.
15
16   Question: What is power in international relations?
17
18   Document [1]: ... the following concepts of political power :
19   - Power as a goal of states or leaders;
20   - Power as a measure of influence or control over outcomes, events, actors and issues;
21   - Power as victory in conflict ...
22   Document [2]: ... American political scientist Joseph Nye Jr., who defined power as "the
         ability to influence the behaviors of others to get the desired outcome." ...
23   ...
24
25   Answer:
```

Figure B.5: Example prompt for the content selection subtask in the LFQA setting.

```
 1   In this task, you are presented with a question and several passages, where some parts are
         "highlighted" (namely, there are <highlight_start> and <highlight_end> tokens before
         and after each such span). Those spans are supposed to have the information needed to
         answer the question, and the goal is to fuse all those highlights into a single
         paragraph that answers the question. As an intermediate step, you need to cluster
         highlights that can be merged into a sentence (namely, each cluster will be later
         merged into one sentence). Make sure the clusters are in the same order you would then
         write the corresponding output paragraph. IMPORTANT: make sure there are at least two
         clusters. IMPORTANT: The output must be of the format [{"cluster":[comma-delimited
         highlights indices]}]
 2
 3   Question: What was the color of the carpet at the Oscars this year, and why wasn't it red?
 4
 5   Document [1]: ... tasked to <highlight_start>reimagine the carpet color<highlight_end>,
         opting to break tradition <highlight_start>so the talent could better transition from
         daytime arrivals to the evening<highlight_end>, according to the Hollywood Reporter
         ...
 6   Document [2]: ... rolling out the red carpet. <highlight_start>The celebrated carpet at
         the Oscars<highlight_end> -- the place where stars show off their most elegant and
         outrageous fashions -- <highlight_start>will be a "champagne" color this
         year.<highlight_end> It will be the first time ...
 7   ...
 8
 9   The highlighted spans are:
10   Document [1]:
11    1. reimagine the carpet color
12    2. so the talent could better transition from daytime arrivals to the evening
13   ...
14   Document [2]:
15    5. The celebrated carpet at the Oscars
16    6. will be a "champagne" color this year.
17   ...
18
19   Answer: The highlighted spans are clustered as follows:
20   [{"cluster":[1,2,5,6]}, {"cluster":[9,11]} ...]
21
22   In this task, you are presented with a question and several passages, where some parts are
         "highlighted" (namely, there are <highlight_start> and <highlight_end> tokens before
         and after each such span). Those spans are supposed to have the information needed to
         answer the question, and the goal is to fuse all those highlights into a single
         paragraph that answers the question. As an intermediate step, you need to cluster
         highlights that can be merged into a sentence (namely, each cluster will be later
         merged into one sentence). Make sure the clusters are in the same order you would then
         write the corresponding output paragraph. IMPORTANT: make sure there are at least two
         clusters. IMPORTANT: The output must be of the format [{"cluster":[comma-delimited
         highlights indices]}]
23
24   Question: What is power in international relations?
25
26   Document [1]: ... <highlight_start>Power as a measure of influence or control over
         outcomes, events, actors and issues;<highlight_end>
27   - <highlight_start>Power as victory in conflict<highlight_end> ...
28   Document [2]: ... American political scientist Joseph Nye Jr., who defined power as
         <highlight_start>"the ability to influence the behaviors of others to get the desired
         outcome."<highlight_end> ...
29   ...
30
31   The highlighted spans are:
32   Document [1]:
33    1. Power as a measure of influence or control over outcomes, events, actors and issues;
34   ...
35   Document [2]:
36    5. "the ability to influence the behaviors of others to get the desired outcome."
37   ...
38   Answer: The highlighted spans are clustered as follows:
```

Figure B.6: Example prompt for the clustering subtask in the LFQA setting.

```
 1  In this task, you are presented with a question and several passages, where some parts are
        "highlighted" (namely, there are <highlight_start> and <highlight_end> tokens before
        and after each such span). You are also presented with a prefix of a paragraph. The
        goal is to generate a paragraph that answers the question, based on the passages. The
        given prefix is what you generated so far. You job is to generate the next sentence in
        the paragraph, that covers all and only the "highlighted" spans. Make sure it connects
        well with the prefix, that it covers all and only the "highlighted" spans, and that it
        properly builds towards a full and coherent response to the question.
 2
 3  Question: What was the color of the carpet at the Oscars this year, and why wasn't it red?
 4
 5  Document [1]: ... interview published Friday."<highlight_start>That's always been
        something that the Oscars has had a problem with ever since it's started because it
        begins so early in the day with the sunshine and the heat<highlight_end>," she added.
        "<highlight_start>But everybody's dressed up for a night event<highlight_end> ...
 6  ...
 7
 8  Prefix: The Oscars red carpet was changed to a champagne color this year, to better
        transition from daytime to evening arrivals.
 9
10  The highlighted spans are:
11  Document [1]:
12   1. That's always been something that the Oscars has had a problem with ever since it's
         started because it begins so early in the day with the sunshine and the heat
13   2. But everybody's dressed up for a night event
14  ...
15  Answer:
16  The next sentence is: This transition was necessary, as the Oscars always began very early
        in the day, when it was hot, but everybody was dressed for a night event.
17
18  In this task, you are presented with a question and several passages, where some parts are
        "highlighted" (namely, there are <highlight_start> and <highlight_end> tokens before
        and after each such span). You are also presented with a prefix of a paragraph. The
        goal is to generate a paragraph that answers the question, based on the passages. The
        given prefix is what you generated so far. You job is to generate the next sentence in
        the paragraph, that covers all and only the "highlighted" spans. Make sure it connects
        well with the prefix, that it covers all and only the "highlighted" spans, and that it
        properly builds towards a full and coherent response to the question.
19
20  Document [1]: ... the following concepts of political power :
21  - Power as a goal of states or leaders;
22  - <highlight_start>Power as a measure of influence or control over outcomes, events,
        actors and issues<highlight_end> ...
23  ...
24
25  Prefix: Power in international relations can be defined as the ability to influence or
        control outcomes, events, actors, and issues.
26
27  The highlighted spans are:
28  Document [1]:
29   1. Power as a measure of influence or control over outcomes, events, actors and issues
30  ...
31  Answer:
32  The next sentence is:
```

Figure B.7: Example prompt for iterative, sentence-by-sentence generation approach in the LFQA setting.

```
 1  In this task, you are presented with a question and several passages, where some parts are
        "highlighted" (namely, there are <highlight_start> and <highlight_end> tokens before
        and after each such span). Those spans are supposed to have the information needed to
        answer the question. Your job is to generate a coherent paragraph that answers the
        question, while covering all and only the "highlighted" spans. Build this paragraph
        iteratively, where each time you cluster highlights to build the next sentence. Make
        sure all highlights are included, and that the output paragraph has at least two
        sentences. IMPORTANT: The clustering steps must be of the form: "Highlights
        comma-delimited highlights indices are combined to form sentence
        sentence_id:sentence_text". Additionally, before generating the final output paragraph,
        say "So the final answer is:".
 2
 3  Question: What was the color of the carpet at the Oscars this year, and why wasn't it red?
 4
 5  Document [1]: ... tasked to <highlight_start>reimagine the carpet color<highlight_end>,
        opting to break tradition <highlight_start>so the talent could better transition from
        daytime arrivals to the evening<highlight_end>, according to ...
 6  Document [2]: ... <highlight_start>The celebrated carpet at the Oscars<highlight_end> --
        the place where stars show off their most elegant and outrageous fashions --
        <highlight_start>will be a "champagne" color this year.<highlight_end> It will be ...
 7  ...
 8
 9  Answer:
10  The highlighted spans are:
11  Document [1]:
12   1. reimagine the carpet color
13   2. so the talent could better transition from daytime arrivals to the evening
14  ...
15  Document [2]:
16   5. The celebrated carpet at the Oscars
17   6. will be a "champagne" color this year.
18  ...
19  The highlighted spans are combined as follows:
20  Highlights 1,2,5,6 are combined to form sentence 1: The Oscars red carpet was changed to
        a champagne color this year, to better transition from daytime to evening arrivals.
21  ...
22  So the final summary is: The Oscars red carpet was changed to a champagne color this year,
        to better transition from daytime to evening arrivals. ...
23
24  In this task, you are presented with a question and several passages, where some parts are
        "highlighted" (namely, there are <highlight_start> and <highlight_end> tokens before
        and after each such span). Those spans are supposed to have the information needed to
        answer the question. Your job is to generate a coherent paragraph that answers the
        question, while covering all and only the "highlighted" spans. Build this paragraph
        iteratively, where each time you cluster highlights to build the next sentence. Make
        sure all highlights are included, and that the output paragraph has at least two
        sentences. IMPORTANT: The clustering steps must be of the form: "Highlights
        comma-delimited highlights indices are combined to form sentence
        sentence_id:sentence_text". Additionally, before generating the final output paragraph,
        say "So the final answer is:".
25
26  Question: What is power in international relations?
27
28  Document [1]: ... <highlight_start>Power as a measure of influence or control over
        outcomes, events, actors and issues;<highlight_end> Power as victory in conflict ...
29  Document [2]: ... Joseph Nye Jr., who defined power as <highlight_start>"the ability to
        influence the behaviors of others to get the desired outcome."<highlight_end> ...
30  ...
31
32  Answer:
33  The highlighted spans are:
34  Document [1]:
35   1. Power as a measure of influence or control over outcomes, events, actors and issues;
36  ...
37  Document [2]:
38   5. "the ability to influence the behaviors of others to get the desired outcome."
39  ...
40  The highlighted spans are combined as follows:
```

Figure B.8: Example prompt for the CoT approach in the LFQA setting.

```
 1  Input:
 2  Those documents were cleared earlier Thursday morning, according to spokesmen for both
        sides. Nevertheless, the drama over the emails once marked ""committee confidential""
        engulfed much of the third day of Kavanaugh's confirmation hearings, with Booker
        vowing to disclose them in defiance of Senate rules.
 3
 4  ""I openly invite and accept the consequences of releasing that email right now,"" Booker
        said. ""
 5
 6  Such theatrics have characterized Kavanaugh's hearings, in which Democrats have
        repeatedly complained that Republicans have withheld documents from the committee and
        the public that shed important light on Kavanaugh's past.
 7  ...
 8  <doc-sep>
 9  New Jersey Democrat Cory Booker said he was ready to risk expulsion from the Senate for
        making public documents pertinent Kavanaugh's time as a top White House aide to
        President George W. Bush, but the GOP mocked him for grandstanding ahead of a
        possible 2020 run -- saying it had already cleared the emails for release.
10
11  In a striking political gambit, Booker, backed up by Democratic Sen. Mazie Hirono of
        Hawaii, released emails -- which had been designated as ""committee confidential"" --
        that reference Kavanaugh's position on racial profiling and thoughts on Roe v. Wade
        dating from his time as a White House official under Bush.
12  ...
13
14
15  Output:
16  the drama over the emails once marked ""committee confidential"" engulfed much of the
        third day of Kavanaugh's confirmation hearings, with Booker vowing to disclose them
        in defiance of Senate rules<highlight_sep>""I openly invite and accept the
        consequences of releasing that email right now,"" Booker said<highlight_sep>Democrats
        have repeatedly complained that Republicans have withheld documents from the
        committee and the public that shed important light on Kavanaugh's
        past...<doc-sep>Booker<highlight_sep>released emails -- which had been designated as
        ""committee confidential"" -- that reference Kavanaugh's position...
```

Figure B.9: Example input and output for the fine-tuned content selection model in the MDS setting. Highlight start and end markups are colored blue.

```
 1  Input:
 2  Those documents were cleared earlier Thursday morning, according to spokesmen for both
         sides. Nevertheless, <highlight_start>the drama over the emails once marked
         ""committee confidential"" engulfed much of the third day of Kavanaugh's confirmation
         hearings, with Booker vowing to disclose them in defiance of Senate
         rules<highlight_end>.
 3
 4  <highlight_start>""I openly invite and accept the consequences of releasing that email
         right now,"" Booker said<highlight_end>. ""
 5
 6  Such theatrics have characterized Kavanaugh's hearings, in which
         <highlight_start>Democrats have repeatedly complained that Republicans have withheld
         documents from the committee and the public that shed important light on Kavanaugh's
         past<highlight_end>.
 7  ...
 8  <doc-sep>
 9  New Jersey Democrat Cory Booker said he was ready to risk expulsion from the Senate for
         making public documents pertinent Kavanaugh's time as a top White House aide to
         President George W. Bush, but the GOP mocked him for grandstanding ahead of a
         possible 2020 run -- saying it had already cleared the emails for release.
10
11  In a striking political gambit, <highlight_start>Booker<highlight_end>, backed up by
         Democratic Sen. Mazie Hirono of Hawaii, <highlight_start>released emails -- which had
         been designated as ""committee confidential"" -- that reference Kavanaugh's
         position<highlight_end> on racial profiling and thoughts on Roe v. Wade dating from
         his time as a White House official under Bush.
12  ...
13
14  Output:
15  the drama over the emails once marked ""committee confidential"" engulfed much of the
         third day of Kavanaugh's confirmation hearings, with Booker vowing to disclose them
         in defiance of Senate rules<doc-sep>Booker<highlight_separator>released emails --
         which had been designated as ""committee confidential"" -- that reference Kavanaugh's
         position<cluster_separator>""I openly invite and accept the consequences of releasing
         that email right now,"" Booker said...
```

Figure B.10: Example input and output for the fine-tuned sentence planning model in the MDS setting. Highlights separators are colored blue and cluster separators are colored green.

```
1  Input:
2  Those documents were cleared earlier Thursday morning, according to spokesmen for both
       sides. Nevertheless, <highlight_start>the drama over the emails once marked
       "committee confidential" engulfed much of the third day of Kavanaugh\'s confirmation
       hearings, with Booker vowing to disclose them in defiance of Senate
       rules<highlight_end>. <doc-sep>New Jersey Democrat Cory Booker said he was ready to
       risk expulsion from the Senate for making public documents pertinent Kavanaugh\'s
       time as a top White House aide to President George W. Bush, but the GOP mocked him
       for grandstanding ahead of a possible 2020 run -- saying it had already cleared the
       emails for release. In a striking political gambit,
       <highlight_start>Booker<highlight_end>, backed up by Democratic Sen. Mazie Hirono of
       Hawaii, <highlight_start>released emails -- which had been designated as "committee
       confidential" -- that reference Kavanaugh\'s position<highlight_end> on racial
       profiling and thoughts on Roe v. Wade dating from his time as a White House official
       under Bush. The New York Times first reported the email. <highlight_start>Throughout
       the contentious hearing, now in its third day, Kavanaugh has tried to give political
       questions --<highlight_end> especially those related to Trump and his potential legal
       woes -- a wide berth. <doc-sep> Senators Dispute Status Of Released Documents As
       <highlight_start>Questioning Of Kavanaugh Wraps Up<highlight_end> Enlarge this image
       toggle caption Chip Somodevilla/Getty Images Chip Somodevilla/Getty Images Last
       night, he was admonished by Republicans for breaking the rules when he read from
       committee confidential documents. Cory and Senate Democrats were able to shame the
       committee into agreeing to make last night\'s documents publicly available, and
       <highlight_start>Cory publicly released those documents as well as other committee
       confidential documents<highlight_end> today. <highlight_start>And he\'ll keep
       releasing them because Republicans are hiding Brett Kavanaugh\'s record from the
       American people<highlight_end>.
3
4  Output:
5  Day three of the Brett Kavanaugh confirmation hearing has opened with a bang: Democratic
       Sen. Cory Booker says he has ordered his staff to release a Kavanaugh email even
       though it's supposed to remain under wraps.
```

Figure B.11: Example input and output for the fine-tuned sentence-by-sentence generation in the MDS setting. This is the first sentence generated in the summary, so there is no prefix.

```
 1
 2   Input:
 3   Q: 'Wild law': if animals have rights, can nature have rights too?
 4   <query_documents_separator>
 5   The fields of animal law and environmental law have an uneasy relationship. At a basic
         level, they are intertwined by the fundamental observation that animals, human and
         nonhuman, exist in the environment. Environmental law is generally concerned with
         animals at the level of species (and specifically endangered or threatened species),
         whereas animal law is concerned with all animals, regardless of particular
         characteristics. The issue of wild horses in the western United States illustrates
         this tension. Some environmentalists view the horses as "feral pests" that damage the
         fragile ecosystem and compete with wildlife - and privately owned cattle - for
         resources.\n
 6   ...
 7   <doc-sep>
 8   Every now and then, a concept emerges with the power to alter orthodoxies. In the legal
         world, one such concept is the "rights of nature," an idea that can-and
         should-encompass new protections for wild animals. Every now and then, a concept
         emerges with the power to alter orthodoxies. In the legal world, one such concept is
         the "rights of nature." Whereas environmental law traditionally begins from the
         premise that humans need a healthy environment for their own purposes ( e.g., to
         enjoy fresh air, clean water, and the many benefits of biodiversity), the rights of
         nature concept flows from the notion that nature is itself entitled to rights
         regardless of human interests. The shift is significant: nature moves from being an
         object of the law to being a subject of the law and a holder of its own rights. In
         contrast to environmental rights, it is only a fairly recent phenomenon that assigns
         "constitutional significance to the experiences of individual nonhuman animals."
 9
10   Output:
11   In contrast to environmental rights, it is only a fairly recent phenomenon that assigns
         "constitutional significance to the experiences of individual nonhuman
         animals."<highlight_sep>At the constitutional level, Ecuador was the first to
         recognize the rights of nature.<highlight_sep>At the most fundamental level, if
         nature has rights, and if nature includes animals, then rights-based claims could be
         made on behalf of animals using existing rights of nature doctrine and
         strategy.<doc-sep>In the legal world, one such concept is the "rights of nature," an
         idea that can-and should-encompass new protections for wild
         animals.<doc-sep>According to the "Rights of Nature" doctrine, an ecosystem is
         entitled to legal personhood status and as such, has the right to defend itself in a
         court of law against harms, including environmental degradation caused by a specific
         development project or even by climate change.<doc-sep>Looking at legislation, the
         recognition of animal sentience is another emerging trend, although this does not go
         so far as to establish rights for animals.<highlight_sep>To assert these rights,
         giving rivers legal personhood helps protect them as they can then be represented in
         court.<highlight_sep>This has occurred with all the rivers in Bangladesh and the
         Atrato River in Columbia, and many others.<doc-sep>In 1972, Supreme Court Justice
         William O. Douglas first asserted the legal rights of nature in a U.S. court in a
         dissenting opinion on the Sierra Club v. Morton case.<highlight_sep>Earlier this
         year, the U.S. District Court for the Southern District of Ohio granted the
         application , thereby officially recognizing animals - hippos - have legal rights
         akin to humans for the first time in U.S. history.
```

Figure B.12: Example input and output for the fine-tuned content selection model in the LFQA setting. Highlight start and end markups are colored blue.

```
 1  Input:
 2  Q: 'Wild law': if animals have rights, can nature have rights too?
 3  <query_documents_separator>
 4  Most constitutions address the environment, and the typical phrasing is anthropocentric:
        a human right to a healthy environment as seen, for example, in article 42 of the
        Constitution of Kenya: "Every person has the right to a clean and healthy environment
        . . . ." Newer trends adopt ecocentric or biocentric approaches and grant rights to
        nature (or its component parts, such as a river) at the constitutional or legislative
        level or through judicial decisions. <highlight_start>In contrast to environmental
        rights, it is only a fairly recent phenomenon that assigns "constitutional
        significance to the experiences of individual nonhuman animals."<highlight_end>
        Ecocentric or biocentric approaches that lodge a right in nature or its component
        parts also may be promising for the development of legally recognized animal rights.
        Rights of nature are not widespread, but they have potential for growth and impact.
        <highlight_start>At the constitutional level, Ecuador was the first to recognize the
        rights of nature.<highlight_end> Bolivia adopted this approach through the Law on the
        Rights of Mother Earth (2010); the enumerated rights are the rights to life,
        diversity of life, water, clean air, equilibrium, restoration, and pollution-free
        living. Other countries have recognized the right in judicial opinions.
        <highlight_start>At the most fundamental level, if nature has rights, and if nature
        includes animals, then rights-based claims could be made on behalf of animals using
        existing rights of nature doctrine and strategy.<highlight_end><doc-sep>Every now and
        then, a concept emerges with the power to alter orthodoxies.\n<highlight_start>In the
        legal world, one such concept is the "rights of nature," an idea that can-and
        should-encompass new protections for wild animals.<highlight_end><doc-sep>How would
        such a legal system work, and could giving rights to nature help in the legal battle
        against climate change? A few case studies offer some insight.
        <highlight_start>According to the "Rights of Nature" doctrine, an ecosystem is
        entitled to legal personhood status and as such, has the right to defend itself in a
        court of law against harms, including environmental degradation caused by a specific
        development project or even by climate change.<highlight_end>
 5
 6  Output:
 7  In contrast to environmental rights, it is only a fairly recent phenomenon that assigns
        "constitutional significance to the experiences of individual nonhuman
        animals."<highlight_separator>At the constitutional level, Ecuador was the first to
        recognize the rights of nature.<doc-sep>Looking at legislation, the recognition of
        animal sentience is another emerging trend, although this does not go so far as to
        establish rights for animals.<highlight_separator>To assert these rights, giving
        rivers legal personhood helps protect them as they can then be represented in
        court.<highlight_separator>...<cluster_separator>At the most fundamental level, if
        nature has rights, and if nature includes animals, then rights-based claims could be
        made on behalf of animals using existing rights of nature doctrine and
        strategy.<cluster_separator>In the legal world, one such concept is the "rights of
        nature," an idea that can-and should-encompass new protections for wild
        animals.<cluster_separator>According to the "Rights of Nature" doctrine, an ecosystem
        is entitled to legal personhood status and as such, has the right to defend itself in
        a court of law against harms, including environmental degradation caused by a
        specific development project or even by climate change.
```

Figure B.13: Example input and output for the fine-tuned sentence planning model in the LFQA setting. Highlights separators are colored blue and cluster separators are colored green.

```
 1  Input:
 2  Q: 'Wild law': if animals have rights, can nature have rights too?
 3  <query_documents_separator>
 4  Most constitutions address the environment, and the typical phrasing is anthropocentric:
        a human right to a healthy environment as seen, for example, in article 42 of the
        Constitution of Kenya: "Every person has the right to a clean and healthy environment
        . . . ." Newer trends adopt ecocentric or biocentric approaches and grant rights to
        nature (or its component parts, such as a river) at the constitutional or legislative
        level or through judicial decisions. <highlight_start>In contrast to environmental
        rights, it is only a fairly recent phenomenon that assigns "constitutional
        significance to the experiences of individual nonhuman animals."<highlight_end>
        Ecocentric or biocentric approaches that lodge a right in nature or its component
        parts also may be promising for the development of legally recognized animal rights.
        Rights of nature are not widespread, but they have potential for growth and impact.
        <highlight_start>At the constitutional level, Ecuador was the first to recognize the
        rights of nature.<highlight_end><doc-sep>Mainly, they have helped in cases of
        mistreatment of chimpanzees and elephants in captivity through petitioning for writs
        of habeas corpus in challenging the detainment of these animals. One of their longest
        running cases concerns Happy , an elephant kept alone in the Bronx Zoo (since 1977)
        who was ruled in June to not be a legal person, though two dissenters stressed
        Happy's rights of liberty and autonomy. <highlight_start>Looking at legislation, the
        recognition of animal sentience is another emerging trend, although this does not go
        so far as to establish rights for animals.<highlight_end> Put into action, a focus is
        on giving ecosystems legal guardianship and personhood. One example of this is with
        rights of rivers. <highlight_start>To assert these rights, giving rivers legal
        personhood helps protect them as they can then be represented in
        court.<highlight_end> <highlight_start>This has occurred with all the rivers in
        Bangladesh and the Atrato River in Columbia, and many others.<highlight_end><doc-sep>
        "I had to choose a single sort of leader in the rights of nature movement; it would
        definitely be indigenous governments in the United States," Bradshaw says. But it
        took many years for these ideas to gain official consideration outside tribal courts
        and Western judicial systems. <highlight_start>In 1972, Supreme Court Justice William
        O. Douglas first asserted the legal rights of nature in a U.S. court in a dissenting
        opinion on the Sierra Club v. Morton case.<highlight_end> Colombian officials planned
        to kill the hippos, but an attorney filed a lawsuit on behalf of them in 2020, hoping
        to save them from euthanasia. Experts in the U.S. applied to provide testimony in the
        Colombian court. <highlight_start>Earlier this year, the U.S. District Court for the
        Southern District of Ohio granted the application , thereby officially recognizing
        animals - hippos - have legal rights akin to humans for the first time in U.S.
        history.<highlight_end>
 5
 6  Output:
 7  The idea of extending legal rights to non-human entities, including animals and nature,
        is not new.
```

Figure B.14: Example input and output for the fine-tuned sentence-by-sentence generation in the LFQA setting. This is the first sentence generated in the summary, so there is no prefix.

Attributable Text Generation Evaluation

In this task you are provided with a question, a generated answer by an AI model and a set of source documents that the model used to answer the question. The generated answer will have citations to snippets of the source documents. Your job is to evaluate the quality of the generated answer and the citations.

```
Query: where does the ohio river start and finish
Response: The Ohio River has its own unique history, as new settlements grew and expanded, sending human and material traffic up
and down the river. It flows southwest joining the Mississippi River near Cairo, which empties into the Gulf of Mexico. However,
the Ohio River is navigable and, despite seasonal fluctuations that occasionally reach flood proportions, its fairly uniform flow
has supported important commerce since settlement first began.
```

The response is fluent and choesive.

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |

The response is helpful and informative answer to the query.

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |

1 Answer Quality ———— 2 Citation Quality ———— 3 Submit

BACK   NEXT

Figure E.1: A screenshot from the first step in our human evaluation annotation interface for the LFQA setting, which is a replication of the interface presented by Liu et al. (2023). The annotator is presented with a query and a response and has to rate two aspects of the response on a Likert scale between 1 to 5. The first aspect is the fluency and cohesiveness of the response. The second is the helpfulness of the response.

Attributable Text Generation Evaluation

In this task you are provided with a question, a generated answer by an AI model and a set of source documents that the model used to answer the question. The generated answer will have citations to snippets of the source documents. Your job is to evaluate the quality of the generated answer and the citations.

```
Query: where does the ohio river start and finish
Response: The Ohio River has its own unique history, as new settlements grew and expanded, sending human and material
traffic up and down the river. It flows southwest joining the Mississippi River near Cairo, which empties into the Gulf
of Mexico. However, the Ohio River is navigable and, despite seasonal fluctuations that occasionally reach flood
proportions, its fairly uniform flow has supported important commerce since settlement first began.
```

✓ The Ohio River has its own unique history, as new settlements grew and expanded, sending human and material traffic up and down the river.

Choice	
Full Support	▼

Do the highlights in the citation(s) `KEYSTONEANSWERS.COM/WHERE-...` fully support the **entire** sentence above?

It flows southwest joining the Mississippi River near Cairo, which empties into the Gulf of Mexico.

Choice	
	▼

Do the highlights in the citation(s) `A-Z-ANIMALS.COM/BLOG/WHERE-DO...` `KEYSTONEANSWERS.COM/WHERE-...` fully support the **entire** sentence above?

```
Enter your email in the box below to get the most mind-blowing animal stories and videos
delivered directly to your inbox every day. – The Ohio River begins at Point State Park in
Pittsburgh, PA, and flows 981 miles to its ending point in Cairo, Illinois. – The Ohio
drainage basin spreads across 14 states and spans 205,000 square miles. – Beyond being
home to the main ports of Cincinnati, Louisville, and Pittsburgh, the Ohio River is a
significant transportation route for the states through which it courses. The Ohio River
flows across six states, and its basin spreads over 14 different states. As with most
waterways, this one also supports many cities and a thriving commercial shipping industry
that contributes significantly to the states' economies through which the river flows.
Besides this, the Ohio River sustains a vast selection of wildlife, with some animals on
endangered and protected lists. So, if the Ohio River and its tributaries flow across 14
states, where does the Ohio River start? The Ohio River is one of the major rivers in the
United States. It flows over 981 miles from Pittsburgh, Pennsylvania to Cairo, Illinois,
where it meets the Mississippi River. This historic river has played a significant role in
the development of the region and continues to be an important resource today. With that
said, here are some cool facts about the Ohio River: – The Ohio River consists of the
Allegheny and Monongahela Rivers in Pittsburgh, Pennsylvania. – The Ohio River flows
```

Figure E.2: A screenshot from the second step in our human evaluation annotation interface for the LFQA setting, which is an adaption of the interface suggested by Liu et al. (2023). The annotator is presented with citations provided by the model. For each sentence, they have to rate the quality of the citations. Highlights are presented if available, with arrows to navigate between different highlights. Completed sentences are presented with a green arrow to indicate that the annotator can continue to the next sentence.