Unstructured Evidence Attribution for Long Context Query Focused Summarization

Anonymous ACL submission

Abstract

Large language models (LLMs) are capable of generating coherent summaries from very long contexts given a user query, and extracting and citing evidence spans helps improve the trustworthiness of these summaries. Whereas previous work has focused on evidence citation with fixed levels of granularity (e.g. sentence, paragraph, document, etc.), we propose to extract *unstructured* (i.e., spans of any length) evidence in order to acquire more relevant and consistent evidence than in the fixed granularity case. We show how existing systems struggle to copy and properly cite unstructured evidence, which also tends to be "lost-in-the-middle". To help models perform this task, we create the **S**ummaries with **Uns**tructured **E**vidence **T**ext dataset (SUnsET), a synthetic dataset generated using a novel pipeline, which can be used as training supervision for unstructured evidence summarization. We demonstrate across 5 LLMs and 4 datasets spanning human written, synthetic, single, and multi-document settings that LLMs adapted with SUnsET generate more relevant and factually consistent evidence with their summaries, extract evidence from more diverse locations in their context, and can generate more relevant and consistent summaries than baselines with no fine-tuning and fixed granularity evidence. We release SUnsET and our generation code to the public.[1]

1 Introduction

At the frontier of the capabilities of natural language processing (NLP) systems such as large language models (LLMs) is the ability to handle long contexts, such as books and sets of research papers, and summarize them based on queries (Koh et al., 2023; Su et al., 2024; Beltagy et al., 2020; Reid et al., 2024). While LLMs have progressed much on this (Edge et al., 2024), people prefer to use traditional retrieval sources (e.g., search engines) for critical queries due to the need for transparency and provenance (Worledge et al., 2024). Citing evidence in the summary addresses this, with prior work first segmenting the context into spans at potentially multiple levels or granularity (e.g., sentences or documents) Li et al. (2023) and having models select evidence from among these segments to support the summary. As has been noted both in work on multi-document summarization (Ernst et al., 2024; Xiao, 2023) and automated fact checking (Wan et al., 2021), this approach is suboptimal for acquiring the most *salient* text in the context to support the summary, resulting in either too much or not enough information. In order to improve the precision of evidence in long-context query focused summarization (LCQFS), we propose to study *unstructured* evidence citation, where any span of arbitrary length within the context can be used as evidence.

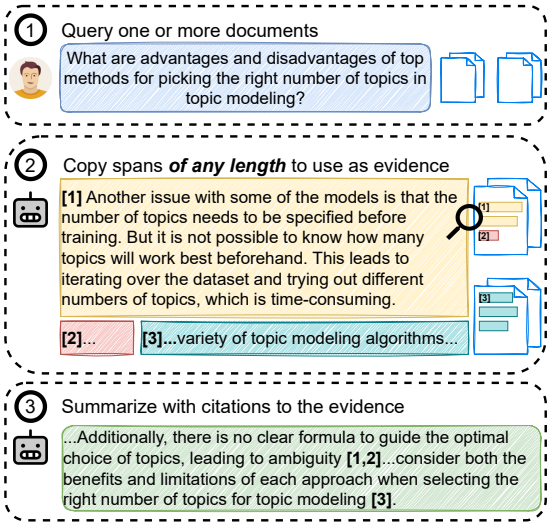


Figure 1: Summarization with *unstructured* evidence requires a model to retrieve spans of any arbitrary length from the context to support individual sentences in the summary. Example given from Llama 3.1 8B trained on our dataset (SUnsET).

[1] https://anonymous.4open.science/r/sunset-BD72/README.md

In the unstructured evidence setup, a model must first copy spans from the context and subsequently use those spans as evidence in the summary (see Figure 1). As we will show, simply prompting LLMs to perform this task with no other intervention leads to poor performance. Thus, we need to adapt models, e.g. through fine-tuning or in-context learning. For this, no suitable training data exist which consists of examples of long documents, queries, summaries, and extracted evidence pointing to arbitrary spans in the documents. Based on the size and cost of other datasets for LCQFS (Asai et al., 2024; Laban et al., 2024; Santosh et al., 2024), this would take an extensive amount of time, money, and expertise to create manually.

To address this, we present a synthetic dataset called the **S**ummaries with **Uns**tructured **E**vidence **T**ext dataset (SUnsET). SUnsET is generated using a novel pipeline, resulting in long documents paired with queries, summaries, and evidence spans. We show that the data in SUnsET are high quality and diverse, comparable to human written data. Using SUnsET, we perform experiments across 5 models and 4 test datasets (including single- and multi-document, human and synthetic data), leading to the following findings: 1) for base LLMs with no fine tuning, extracting and citing unstructured evidence is challenging, and evidence is often lost-in-the-middle; 2) training on documents with shuffled structure (facilitated by SUnsET) can help mitigate lost-in-the-middle, and 3) learning to cite unstructured evidence improves citation accuracy and coverage over fixed-granularity evidence, and additionally improves summary quality.

In sum, our contributions are:

- A synthetic dataset (SUnsET) generated using a novel pipeline
- The first study on unstructured evidence citation for LCQFS, demonstrating that models adapted with SUnsET produce higher quality evidence and summaries than baselines
- An analysis of and method to reduce the lost-in-the-middle problem with unstructued evidence

2 Challenges in LCQFS

LCQFS requires a model to be able to simultaneously ingest a large number of context tokens (possibly from multiple documents), retrieve and attend to relevant information in this context given a query, and integrate this information into a factually consistent and relevant summary. LLMs, with their increasingly large context sizes, have proven to be particularly adept at performing this task (Zhang et al., 2024a; Edge et al., 2024; Russak et al., 2024). Yet, a number of challenges remain, both in dealing with long contexts and with producing query-focused summaries (Li et al., 2024; Russak et al., 2024; Bai et al., 2024; Liu et al., 2024c; Shaham et al., 2023; Ravaut et al., 2024; Laban et al., 2024; Worledge et al., 2024; Ji et al., 2023; Ernst et al., 2024). The main foci of our work are evidence attribution (Laban et al., 2024; Worledge et al., 2024; Li et al., 2023; Ernst et al., 2024; Fierro et al., 2024) and evidence being lost-in-the-middle (Liu et al., 2024c; Ravaut et al., 2024), described next.

Fixed-Granular Single Sentence Citation:
SUMMARY SNIPPET: ...[48] explains that the legend of the Ghost Ship is often told by space men as a cautionary tale....

EVIDENCE: [48] He had heard it spoken of in whispers by drunken space men and professional tellers of fairy tales.

Unstructured Citation:
SUMMARY SNIPPET: ...he, like the ship's former crew, is doomed to wander in space, never able to return to Earth, a haunting reminder of what he has lost and what he can never have [2]...

EVIDENCE: [2] Doomed for all eternity to wander in the empty star-lanes, the Ghost Ship haunts the Solar System that gave it birth. And this is its tragedy, for it is the home of spacemen who can never go home again.

Figure 2: Examples of fixed-granular and unstructured evidence generated by models in our study. Fixed granular citations may include irrelevant or not enough information to support their citing sentences. Unstructured evidence allows for more flexible and precise evidence.

2.1 Evidence Attribution

Improving the ability of LLMs to generate both relevant summaries and provide accurate attributions has the potential to help improve their usefulness, transparency, and trustworthiness. Recent work has started to explore this direction for LCQFS, including SummHay (Laban et al., 2024) and OpenScholar (Asai et al., 2024). However, most works focus on fixed-granularity evidence (e.g., spans, sentences, paragraphs, or documents, Li et al. (2023)). Being able to flexibly cite evidence of any arbitrary length can lead to higher quality summaries which use *precise* pieces

of evidence from the context (Wan et al., 2021; Ernst et al., 2024; Xiao, 2023), as opposed to full documents which contain irrelevant information or individual sentences which may contain not enough information (see e.g., Figure 2). To the best of our knowledge, we provide a first study on unstructured evidence citation in LCQFS with LLMs.

2.2 Lost-in-the-Middle

LLMs suffer from positional preferences in their learned attention (Liu et al., 2024c), oftentimes preferring early or late tokens in their context (Zhang et al., 2024b). While this problem was originally demonstrated on retrieval-augmented-generation (RAG) tasks with explicit answers such as question answering, follow-up work has shown its persistence in more abstractive tasks such as summarization (Ravaut et al., 2024) and query focused multi-document summarization (Laban et al., 2024). A number of solutions have been proposed, most of which rely on manipulating either the positions of tokens in the context or the positional embeddings of LLMs in order to remove their intrinsic bias (Wang et al., 2025; He et al., 2024; Zhang et al., 2024b). We explore and document this problem at the level of unstructured evidence citation, demonstrating how evidence is extracted unevenly across documents, and how this problem can be mitigated using purely synthetic data.

3 Learning to Use Unstructured Evidence

Our task is: given a query about a long input consisting of one or more documents, generate a response to the query which cites arbitrary length text spans from the input. This introduces challenges over the fixed-granularity case (Laban et al., 2024; Asai et al., 2024; Li et al., 2023), as targeted, precise evidence spans must be accurately copied from the context which are relevant and consistent with the summary sentences. While challenging, this can lead to summaries with more accurate and supportive evidence (Ernst et al. 2024).

Large scale synthetic datasets are useful for fine-tuning task specific models at a lower cost than manual annotation (Ziegler et al., 2024; Honovich et al., 2023; Wang et al., 2023; Chen et al., 2024; Xu et al., 2024). To train LLMs to use unstructured evidence, we create SUnsET, a synthetic dataset based on a novel inductive generation pipeline. Training is performed using adapters (Houlsby et al., 2019) to improve unstructured evidence citation and mitigate the lost in the middle problem. For the latter, previous work has shown that fine-tuning with data augmentation (e.g., shuffling documents; Zhang et al., 2024b) can help achieve this. Given this, we construct SUnsET so that documents are modular: documents are broken down into discrete sections, so that data augmentation through shuffling document sections (thus shuffling global structure) is possible. We first present the inductive pipeline approach used to generate SUnsET, followed by our two fine-tuning schemes.

P1. Titles: Generate N unique titles of fiction and non-fiction documents.
P2. Document outline: Given a title, generate an outline broken down into discrete sections.
P3. Queries, summaries, and evidence: Given a document title and outline, generate 5 questions, 5 responses, and supporting passages that will be included in the document.
P4. Document sections: Generate each section of the document one at a time. Ensure that evidence passages are included verbatim.
P5. Refinement: For each ⟨question, summary, evidence⟩ tuple, refine the summary and evidence based on the document.
P6. Validation: For each ⟨question, summary, evidence⟩ tuple, validate that the summary fully addresses the question, is faithful to the document, and includes inline attribution to evidence passages.

Figure 3: Six stage inductive data generation pipeline. The full prompts for each stage are given in Appendix A Figure 8 - Figure 16.

3.1 Generating SUnsET

Our pipeline generates long documents paired with queries, and summaries which address those queries. Each summary additionally includes citations which reference relevant text spans in the original document. We make several design decisions intended to overcome known problems in synthetic data generation, including the potential for low diversity (Honovich et al., 2023; Wang et al., 2023) and labeling errors (Chen et al., 2024). This includes taking a six stage pipeline approach which generates synthetic data inductively, and validation steps which refine summaries, refine evidence, and reject bad summaries and evidence.

The full generation process is described in Figure 3, with prompts provided in Appendix A. Diversity in document topic and type is accomplished by first generating diverse document titles, which seed the subsequent steps of generation. We inductively

	SUnsET			Non-Pipelined			Title + Doc		
Metric	Q	S	D	Q	S	D	Q	S	D
TTR	**0.75**	**0.84**	**0.82**	0.67	0.80	0.35	0.63	0.78	0.35
Cos	**0.81**	**0.73**	**0.68**	0.73	0.72	0.04	0.66	0.61	0.04
Len	**13.45**	**226.5**	**3767.4**	9.85	23.79	474.8	10.21	24.45	433.8

Table 1: Statistics and diversity metrics of synthetic data. Metrics are average type-token ratio (TTR) Bestgen (2023), embedding cosine distance (Cos), and average word length (Len). Columns differentiate between (Q)uestion, (S)ummary and (D)ocument metrics in each dataset. **Bold** is highest diversity across datasets.

Dataset	Topic Diversity
Non-Pipelined	0.506
Title + Doc	0.356
SQuALITY (human, stories)	0.705
LexAbSumm (human, legal text)	0.673
ScholarQABench (human, scientific docs)	0.695
SUnsET	0.679

Table 2: Topic diversity scores using the approach from Terragni et al. (2021). Shading indicates magnitude of diversity score.

build up each document, starting with the queries, summaries, and evidence passages. When generating evidence, each evidence passage is assigned to a section in the document so that evidence can be distributed precisely. The summaries, queries, and assigned evidence are then used as context from which each section of the document is generated, one section at a time. This makes documents modular, which we take advantage of during training to study lost-in-the-middle. Following this, the queries, summaries, and evidence are refined by using the final document as context. Finally, we filter out poor summaries and evidence by prompting to predict if the summaries fully address the query and are fully supported by the document (see Figure 22 in Appendix B for an example). In total we generate 2,352 synthetic documents, giving us 11,309 ⟨document, question, summary⟩ tuples. The cost of the pipeline is relatively cheap, and $\sim$1000$\times$ cheaper than for a manually curated dataset (see Appendix E for an analysis).

We evaluate both the quality and diversity of data generated using this pipeline. For quality, we asked two independent annotators (NLP researchers unaffiliated with the project) three questions for 100 ⟨question, summary, evidence⟩ tuples: Q1) Does the summary address the question?; Q2) Is the summary well structured and organized; and Q3) Does the evidence fully support the summary? Annotators responded to each question with one of the following values: 1 - Not at all; 2 - Somewhat; 3 - Completely. We find that the data is very high quality, acquiring scores of 2.99 for Q1, 2.97 for Q2, and 2.90 for Q3, with an exact agreement rate of 93.67% across all 300 annotations.

To validate SUnsET diversity, we generate two baseline datasets. The first is generated by combining all the steps in Figure 3 into one prompt, forcing the model to simultaneously perform all tasks to generate each example (called Non-Pipelined). The second includes a title generation step to seed each document (called Title + Doc, see Figure 17 in Appendix A for prompts). We compare each dataset using samples of 100 documents along lexical and semantic diversity metrics in Table 1. Further, in Table 2 we compare the topic diversity (following Terragni et al. 2021) between these datasets, as well as three human-written datasets: **SQuALITY** (Wang et al., 2022), **LexAbSumm** (Santosh et al., 2024), and **ScholarQABench** (Asai et al., 2024), (see Appendix D. Our approach generates longer documents with longer summaries than baseline non-pipelined approaches, which also tend to be much more diverse. Additionally, our pipeline produces documents with topic diversity similar to that of human written datasets.

3.2 Training Complementary Adapters

Previous work has demonstrated that altering the position embeddings of LLMs either directly or through fine-tuning can help to overcome positional biases (Hsieh et al., 2024; Zhang et al., 2024b). We design SUnsET documents so that they are modular, having global coherence at the level of the full document and local coherence at the level of discrete sections. Given this, we experiment with position-aware and position-agnostic training in order to observe their impact on evidence selection and quality, as well as summary quality. For position-aware training, we concatenate all document sections together in their natural order to construct the context, while for position-agnostic training, we shuffle the document sections before concatenating them, thus randomizing the global structure of the position embeddings while maintaining the local structure. This gives us two adapters for each model in our experiments. The prompt we use for training is provided in Appendix A Figure 18, and all training is performed using supervised fine-tuning on SUnsET data using LoRA (Hu et al., 2022). In all cases we fine tune using the Hugging-

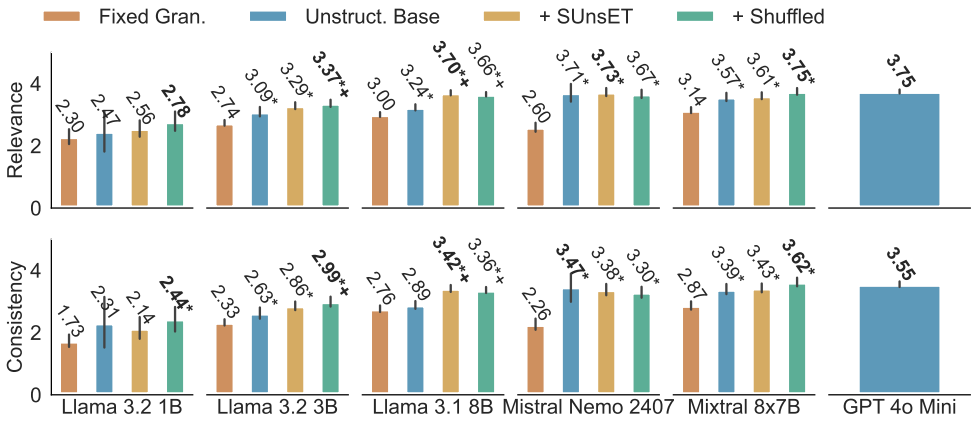


Figure 4: Average relevance and consistency of evidence texts with respect to their citation sentences measured using an autorater (DeepSeek-V3; Liu et al., 2023) based on prompts which have previously undergone human evaluation for quality (Liu et al., 2024b). **Bold** indicates best performance for a given model; "*" and "+" indicate statistical significance above the fixed granularity and non-fine-tuned unstructured baselines, respectively, based on non-overlapping 95% confidence intervals.

Model	Exact Match	50% Match	# Evidence
Llama 3.2 1B	0.0	35.71	14
+ SUnsET	**7.69**	**43.26**	208
+ Shuffle	5.15	22.68	97
Llama 3.2 3B	25.57	**90.11**	1345
+ SUnsET	**52.77**	85.62	3720
+ Shuffle	32.99	74.07	2337
Llama 3.1 8B	43.93	83.12	3412
+ SUnsET	**78.36**	**97.21**	4690
+ Shuffle	54.53	88.51	4684
Mistral Nemo 2407	5.48	66.13	310
+ SUnsET	**82.20**	**97.29**	2107
+ Shuffle	72.38	95.76	1959
Mixtral 8x7B	5.79	91.25	3452
+ SUnsET	**33.82**	90.47	4208
+ Shuffle	29.29	**90.74**	4288
GPT-4o-mini	*11.06*	*96.32*	*8159*

Table 3: Evidence copy rates. We measure exact string match (i.e. when the evidence sentence *exactly* appears in the context) as well as 50% overlap between the extracted evidence and the longest common substring.

face Transformers implementation of LoRA (Hu et al., 2022) with a rank and α of 16 applied to all linear operators of each model.

3.3 Summarizing with Unstructured Evidence

To generate summaries with unstructured evidence, we use the prompt from Asai et al. (2024), altering it to include unstructured evidence extraction as a first step. The full prompt is given in Figure 18 in Appendix A. We use this prompt for both inference and supervised fine-tuning on SUnsET. To deal with long contexts, we divide-and-conquer by chunking each document by the model's maximum token length, summarize each chunk, and finally summarize the summaries. Thus, the output for each ⟨document, query⟩ pair is a ⟨summary, evidence_list⟩ pair containing the summary and a list of evidence text from the context.

4 Experiments and Results

Our experiments focus on three research questions:

- **RQ1:** How well can LLMs extract and use unstructured evidence?
- **RQ2:** Is evidence lost-in-the-middle?
- **RQ3:** Does learning to cite unstructured evidence improve summary quality?

Test Data We use four test datasets (full dataset descriptions in Appendix C). Importantly, these include three human written datasets, forcing models trained on SUnsET to be able to generalize beyond synthetic data. At a high level these are: **SQuALITY** (Wang et al. 2022, short sci-fi novels, single document, average context length: 5,200 tokens); **LexAbSumm** (Santosh et al. 2024, long legal documents, single document, average context length: 14,357 tokens); **SummHay** (Laban et al. 2024, synthetic conversations and news, multi-document, average haystack context length: 93,000 tokens); and **ScholarQABench** (Asai et al. 2024, Computer Science research papers, multi-document, average context length: 16,341 tokens). We present here the average results across datasets, results on individual datasets are presented in Appendix F.

Models We use a set of LLMs covering multiple sizes and pretraining configurations. This includes

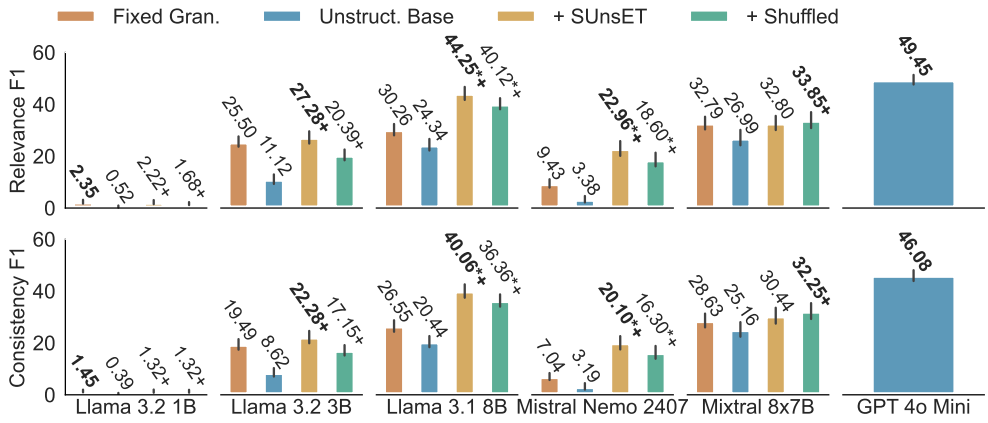


Figure 5: Relevance and consistency F1 scores. **Bold** best performance for a given model; "*" and "+" indicate statistical significance above the fixed granularity and non-fine-tuned unstructured baselines, respectively, based on non-overlapping 95% confidence intervals.

Llama 3.2 1B, Llama 3.2 3B, Llama 3.1 8B (Dubey et al., 2024), Mistral Nemo 2407, and Mixtral 8x7B.[2] We compare four settings for each LLM: base models with fixed granularity evidence (Fixed Gran.), base models with unstructured evidence citation (Unstruct. Base), training adapters on SUnsET (+ SunSET), and training adapters on shuffled SUnsET documents (+ Shuffled). Additionally, we provide an upper bound estimate on performance using GPT 4o mini with no fine-tuning.

Evaluation We follow recent trends in summarization evaluation, which have noted that traditional lexical based metrics such as ROUGE score (Lin, 2004) are insufficient for more complex summarization tasks (Koh et al., 2022). We evaluate our models using autoraters (i.e., LLM-as-a-judge, Gu et al. (2024); Zheng et al. (2023); Liu et al. (2023)) along two dimensions. These dimensions are *Relevance* and *Consistency*. Given a source text, a target text, and optionally a query, *Relevance* measures how well the target covers the main points of the source, as well as how much irrelevant or redundant information it contains. *Consistency* measures to what degree the target contains any factual errors with respect to the source. Both scores are measured on a scale from 1-5 using DeepSeek-V3 (Liu et al., 2024a).[3] We use prompts which have been previously validated to correlate well with human annotations of relevance and consistancy (listed in Appendix A Figure 20 and Figure 21) (Liu et al., 2024b).

[2]Huggingface model IDs are listed in Appendix I Table 8

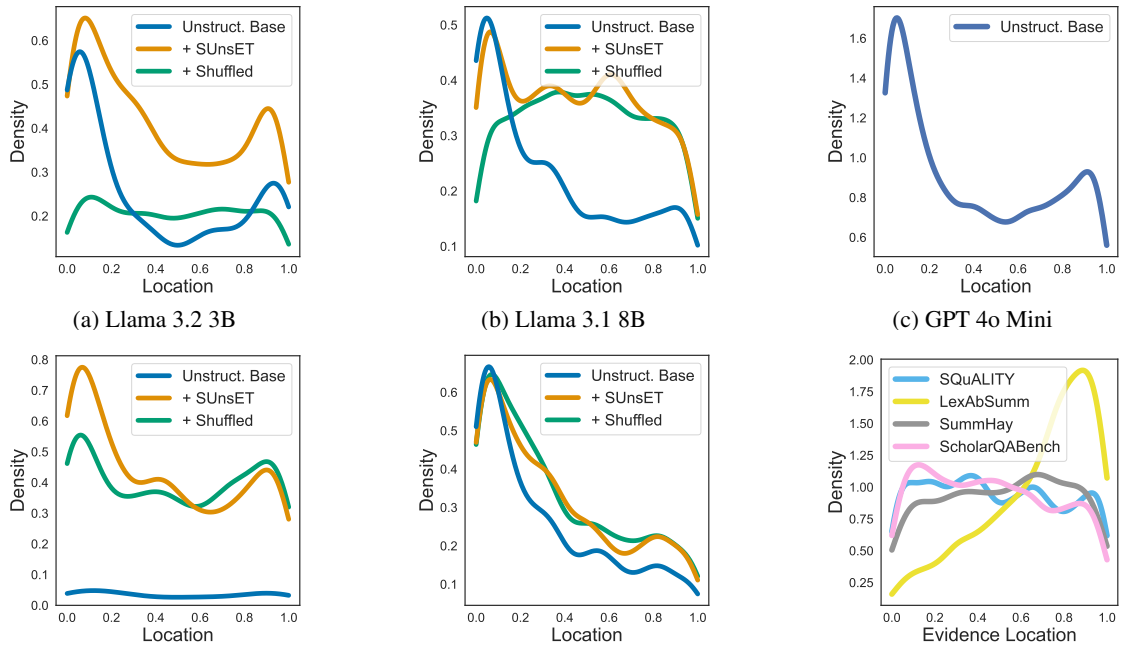
[3]We validate the robustness of the ratings from DeepSeek-V3 in Appendix K.

4.1 RQ1: Can LLMs Use Unstructured Evidence?

Using the datasets and models just described, we first test how well models can copy and utilize unstructued evidence (i.e., any span of arbitrary length from the context). We look at two aspects: evidence copy accuracy, and evidence quality.

Copy Accuracy To study copy accuracy, we match each piece of evidence to its longest common substring (LCS) in the context. We present the rate of exact evidence match and 50% LCS overlap for all models aggregated across all datasets in Table 3. We see that **without fine-tuning, models struggle to copy evidence from the context**. This includes GPT 4o mini, which only copies perfectly 11% of the time. **SUnsET helps models learn to copy evidence spans** in all cases except for the smallest model (Llama 3.2 1B). We see that the number of citations also dramatically increases.

Evidence Quality Next, we measure evidence quality based on the relevance and consistency of evidence spans with their citing sentences using the LLM-as-a-judge setup previously mentioned. We look at two aspects: first, the average citation quality (Figure 4) and second, the citation F1 score (Figure 5), which balances citation quality with the total number of sentences that contain a citation. We calculate the latter similarly to Asai et al. (2024): for a given ⟨summary, evidence_list⟩ pair, we extract all citations from each sentence and normalize their relevance and consistency scores to lie between 0 and 100. For precision, we average these scores over the total number of citations, and for



(a) Llama 3.2 3B (b) Llama 3.1 8B (c) GPT 4o Mini

(d) Mistral Nemo 2407 (e) Mixtral 8x7B (f) Test Datasets

Figure 6: Distribution of location of extracted evidence in the provided source context for different methods. Test dataset evidence location is measured by comparing to reference summaries.

recall, we average the scores over the total number of sentences in the summary.

We find that **the average citation quality of unstructured evidence is better than fixed granularity evidence** (Figure 4). This validates the unstructured evidence approach, where flexible evidence extraction enables higher quality citations to source texts. We also see that models' ability to extract quality evidence is improved by SUnsET, where our results are on par with GPT 4o Mini. When balancing citation quality and citation quantity (Figure 5), we see that **learning to use unstructured evidence with SUnsET leads to statistically significant improvements over fixed-granularity and non-fine-tuned baselines across models.** This is particularly the case for medium to larger models. For smaller models (particularly, Llama 3.1 1B), simply fine-tuning for such a complex task is insufficient, where all settings struggle to extract and use evidence. Non-shuffled training is often better than shuffled training, though shuffled training also improves citation quality by a large margin. When balancing for recall, fixed-granularity evidence tends to be better than unstructured evidence without fine-tuning, which makes sense as a model only needs to generate *references* in the fixed-granularity case. Thus, the primary benefits to citation quality by learning from SUnsET are two-fold: the quality of the evidence itself improves, and the rate of citation improves.

4.2 RQ2: Is evidence lost-in-the-middle?

Next, we quantify to what extent unstructured evidence is lost in the middle. For this, we match extracted evidence to its relative location in the document context (based on 50% LCS overlap) and plot the distributions in Figure 6. As a point of reference, we also plot the distribution of summary sentence locations within the test set documents by matching ground truth reference summaries to their relative locations in their context documents.[4]

We find that **evidence is lost in the middle for all non-fine-tuned models, most often appearing at the beginning or end of the context**. This includes GPT 4o Mini, which has a sharp spike of evidence in the early context. This stands in contrast to ground truth summary location distributions, which are uniform in all cases except for LexAbSumm which has a bias for evidence at the end of the context. In general, training on SUnsET without shuffling increases the rate of evidence extraction, and can help decrease the bias. Shuffling on the other hand, increases the rate of evidence extraction and decreases the bias in all cases ex-

[4]We find the relative location using cosine similarity of S-BERT sentence embeddings (Reimers and Gurevych, 2022)

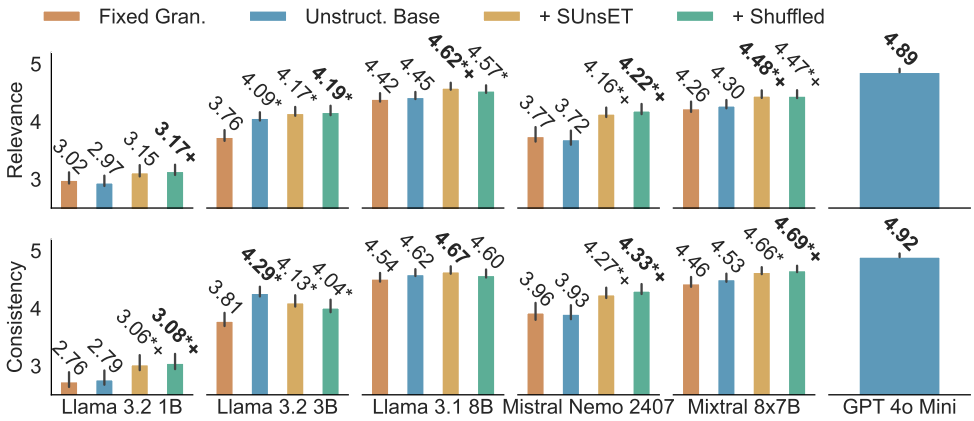


Figure 7: Relevance and consistency of generated summaries. **Bold** best performance for a given model; "*" and "+" indicate statistical significance above the fixed granularity and non-fine-tuned unstructured baselines, respectively, based on non-overlapping 95% confidence intervals.

cept for Mixtral 8x7B. Thus, the modular nature of SUnsET documents, where global structure can be shuffled while local structure is maintained, can be utilized to help reduce positional biases in evidence selection, better reflecting the natural distribution of evidence based on reference data.

4.3 RQ3: Is Summary Quality Improved?

Finally, we test if using unstructured evidence has a positive impact on summary quality. To do so, we measure the relevance and consistency of every summary with respect to its context and query. Our results are presented in Figure 7 (results on individual datasets are given in Appendix F).

First, **for fixed granularity evidence the summaries tend to be similar or slightly lower in quality than unstructured with no fine-tuning, further motivating the unstructured approach.** This is likely because the unstructured evidence task has two subtasks: salient evidence selection, followed by summarization, which has been linked to improvements in summary quality (Ernst et al., 2024). Second, we find that **training on SUnsET leads to statistically significant improvements in summary quality over both baselines.** Standard and shuffled training on SUnsET generally lead to similar gains in performance over unstructured with no fine-tuning, meaning the selection of which approach comes down to a tradeoff between overall evidence quality (where standard has a slight edge) and evidence diversity (where shuffled has an edge). To observe the effect of number of training samples from SUnsET, we perform an ablation where we fine-tune on different number of samples in Appendix G Figure 23 and Figure 24, finding that best performance only requires around 3k samples. Third, **by measuring Pearson's R correlation between citation and summary scores, we find a moderate correlation (0.35 for Relevance and 0.34 for Consistency), demonstrating a relationship between the quality of the citations and the quality of the summaries**. Ultimately, we show the unstructured evidence setup can lead to better evidence and summaries, and demonstrate the utility of SUnsET for learning the task across diverse, human written data.

5 Discussion and Conclusion

Citing precise evidence spans of any arbitrary length for LCQFS has the potential to improve user trust in LLM summaries, as well as the quality of the evidence. Our study highlights salient challenges in this task, contrasts it with the fixed-granular approach, and demonstrates an effective method towards solving it. With no intervention, evidence is lost-in-the-middle, which we show across many settings for the case of unstructured evidence. They additionally struggle to accurately copy arbitrary length evidence from their contexts by default. Our proposed dataset, SUnsET, serves as a useful and inexpensive synthetic dataset to mitigate these issues. This intervention is at training time, meaning the inference cost is lower than for complex reasoning and inference chains. In addition to improving evidence quality, overall summary quality is improved. We hope this work can be built upon to help create more reliable, trustworthy, and useful summarization systems.

Limitations

While our approach offers several benefits, there are notable areas to improve upon. Generating unstructured evidence directly can be prone to hallucination, while it is critical for the evidence to be exactly correct. A more precise RAG approach may offer some benefits. While shuffling during training helps the model to pull evidence more evenly, this also reduces the benefits in terms of evidence quality. A more targeted approach based on directly altering positional embeddings may be more appropriate for this (Hsieh et al., 2024). We experiment with documents using a fixed number of sections in this study; allowing for variable-length documents could deliver greater improvements in performance. Additionally, we acknowledge potential prompt bias influencing model outputs, and that synthetic data may have characteristics which differ from human-written texts. Despite our efforts to mitigate these effects, they persist as a challenge, and using techniques such as APO (Pryzant et al., 2023) could address these issues. Finally, while SUnsET data is domain agnostic, it could be worth exploring how domain-aware data could help for more targeted applications (e.g., in the legal domain).

Ethical Implications

LLMs are capable of generating convincing summaries from long contexts, and learning to generate unstructured supporting evidence from the source context can help improve their reliability and transparency. This approach is more flexible than the fixed-granularity approach, but generation will likely always be prone to errors. Validating that generated evidence is authentic is then crucial, as an incorrect citation presented as a ground truth fact could potentially be more harmful than no citation at all.

Additionally, synthetic data is clearly useful for learning to cite unstructured evidence. But synthetic data comes with its own ethical issues, including plagiarism and copyright infringement. More work on LLM trust and safety is needed to effectively mitigate this, as we are benefitting technologically from unknowing people's free labor.

References

Akari Asai, Jacqueline He*, Rulin Shao*, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Tian, D'arcy Mike, David Wadden, Matt Latzke, Minyang, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Dan Weld, Graham Neubig, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. OpenScholar: Synthesizing Scientific Literature with Retrieval-Augmented Language Models. *CoRR*, abs/2411.14199.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150.

Yves Bestgen. 2023. Measuring lexical diversity in texts: The twofold length problem. *CoRR*, abs/2307.04626.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. AlpaGasus: Training a Better Alpaca with Fewer Data. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock,

Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *CoRR*, abs/2404.16130.

Ori Ernst, Ori Shapira, Aviv Slobodkin, Sharon Adar, Mohit Bansal, Jacob Goldberger, Ran Levy, and Ido Dagan. 2024. The Power of Summary-Source Alignments. In *Findings of the Association for Computational Linguistics, ACL*, pages 6527–6548. Association for Computational Linguistics.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to Plan and Generate Text with Citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11397–11417. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-Judge. *CoRR*, abs/2411.15594.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never Lost in the Middle: Mastering Long-Context Question Answering with Position-Agnostic Decompositional Training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational (ACL)*. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, volume 97, pages 2790–2799.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James R. Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In *Findings of the Association for Computational Linguistics (ACL)*, pages 14982–14995. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR*. OpenReview.net.

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

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2023. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *ACM Comput. Surv.*, 55(8):154:1–154:35.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How Far are We from Robust Long Abstractive Summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A Survey of Large Language Models Attribution. *CoRR*, abs/2311.03731.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can Long-Context Language Models Understand Long Contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. DeepSeek-V3 Technical Report. *CoRR*, abs/2412.19437.

Gabrielle Kaili-May Liu, Bowen Shi, Avi Caciularu, Idan Szpektor, and Arman Cohan. 2024b. MDCure: A Scalable Pipeline for Multi-Document Instruction-Following. *CoRR*, abs/2410.23463.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024c. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7957–7968. Association for Computational Linguistics.

Mathieu Ravaut, Aixin Sun, Nancy F. Chen, and Shafiq Joty. 2024. On Context Utilization in Summarization with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational (ACL)*. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Nils Reimers and Iryna Gurevych. 2022. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Melisa Russak, Umar Jamil, Christopher Bryant, Kiran Kamble, Axel Magnuson, Mateusz Russak, and Waseem AlShikh. 2024. Writing in the Margins: Better Inference Pattern for Long Context Retrieval. *CoRR*, abs/2408.14906.

T. Y. S. S. Santosh, Mahmoud Aly, and Matthias Grabmair. 2024. LexAbSumm: Aspect-based summarization of legal decisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10422–10431. ELRA and ICCL.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021*, pages 263–270. Association for Computational Linguistics.

Hai Wan, Haicheng Chen, Jianfeng Du, Weilin Luo, and Rongzhen Ye. 2021. A DQN-based Approach to Finding Precise Evidences for Fact Verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (ACL/IJCNLP)*, pages 1030–1039. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a Long-Document Summarization Dataset the Hard Way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2025. Eliminating Position Bias of Language Models: A Mechanistic Approach. *CoRR*.

Theodora Worledge, Tatsunori Hashimoto, and Carlos Guestrin. 2024. The Extractive-Abstractive Spectrum: Uncovering Verifiability Trade-offs in LLM Generations. *CoRR*, abs/2411.17375.

Min Xiao. 2023. Multi-doc Hybrid Summarization via Salient Representation Learning. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track (ACL)*, pages 379–389. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2024a. Benchmarking Large Language Models for News Summarization. *Trans. Assoc. Comput. Linguistics*, 12:39–57.

Zheng Zhang, Fan Yang, Ziyan Jiang, Zheng Chen, Zhengyang Zhao, Chengyuan Ma, Liang Zhao, and Yang Liu. 2024b. Position-Aware Parameter Efficient Fine-Tuning Approach for Reducing Positional Bias in LLMs. *CoRR*, abs/2404.01430.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. 2024. CRAFT Your Dataset: Task-Specific Synthetic Dataset Generation Through Corpus Retrieval and Augmentation. *CoRR*, abs/2409.02098.

A List of Prompts

The full set of prompts used in this study are listed in the figures below.

A.1 Synthetic Data Generation Prompts

The prompts used to generated synthetic data are given in Figure 8 – Figure 16.

A.2 Training and Inference Prompt

The prompt used for training and inference is given in Figure 18

A.3 Evaluation Prompts

The prompt used to measure relevance is given in Figure 20 and the prompt used to measure consistency is given in Figure 21.

B SUnsET Example

An example snippet from SUnsET is given in Figure 22

C Full Dataset Descriptions

The test datasets we use in this study include:

SQuALITY (Wang et al., 2022) is a single-document task created from public domain short sci-fi stories where expert annotators create original summaries, providing both an overall narrative and detailed responses to specific questions, challenging models to capture broad context as well as fine-grained information.

LexAbSumm (Santosh et al., 2024) is a single-document task which contains legal judgments from the European Court of Human Rights, focusing on aspect-specific summaries that distill complex legal arguments.

SummHay (Laban et al., 2024) is a multi-document task composed of large-scale "haystacks" of documents with embedded "insights" which are relevant to the queries.

ScholarQABench (Asai et al., 2024) is a multi-document task focused on scientific literature, comprising expert-crafted queries and extended answers drawn from a broad corpus of open-access research papers.

D Topic Diversity Comparison

We have measured the topic diversity of SUnsET using the topic diversity approach from (Terragni et al., 2021). This uses LDA to identify 200 topics across each document, sums up the number of unique words in the first 200 words of each topic, and averages this over a maximum of 200 words * 200 topics (so the score is 1 if each topic has at least 200 unique words, see `https://github.com/MIND-Lab/OCTIS`). We compare this to the two baseline datasets, as well as the human test data, finding that the data in SUnsET is indeed diverse and comparable to human data.

E SUnsET Cost Comparison

Manually annotating data of the kind in SUnsET is highly expensive, requiring annotators to read long sets of documents with long summaries and verifying the quality of the references. As a comparison, SQuALITY (Wang et al., 2022) is a similar dataset to ours in terms of document and response size, and they paid Upwork workers $13 to write each response, followed by $8 to review each response in their data. As we generated 11,309 responses in SUnsET, this alone would have cost $237,468. In contrast, generating SUnsET, including documents, questions, responses, and evidence, cost around $200 (over 1000x cheaper).

P1: Title Generation
Imagine that you must write a book. This book can be either fiction or non-fiction.
You can select any subject to write your book about. Please make the book interesting.
Please write a list of 100 possible book titles.
Please only generate the title for each book.
Please include a mix of fiction and non-fiction, and please try to cover as many genres as possible.
Please make each book title unique.
Please make the style of each book title as different as possible, and don't repeat title styles.
Please generate titles for books which will have a broad range of appeal.
Please generate titles for books which will require a broad range of reading levels.
Please try to make each title as different as possible.
Please do not include many titles with a colon (:).
`{prev_titles_prompt}`

OUTPUT FORMAT

Please separate each book title with a newline character ("\n")

Figure 8: Title generation prompt. `{prev_titles_prompt}` is filled with prompts of previously generated titles.

F Results on Individual Datasets

Results on individual datasets are given in Table 4 (citation precision), Table 5 (citation recall), and Table 6 (F1 score based on citation precision and recall). We see that citation precision is almost uniformly improved across datasets when using unstructured evidence. In other words, when evidence is used within a summary, the evidence is higher quality than fixed granularity evidence in all but 3 cases. This quality is generally further improved by learning from SUnsET. Recall is also improved by learning from SUnsET, and is often better than fixed granularity evidence where a model simply needs to generate reference numbers (as opposed to unstructured where the evidence must also be copied, making the task more challenging). For Llama 3.1 8B and Nemo, overall F1 score is better across all datasets, while for Mixtral and the smaller Llama models the results are mixed across datasets. This is generally because the recall of the fixed granular case tends to be slightly higher, despite referencing lower quality evidences on average. However, when looking at the averages across datasets (Figure 5), we see that learning to cite unstructured evidence with SUnsET leads to the best overall performance.

For summary quality (Table 7), unstructured evidence leads to the best summaries across models and datasets most often, including the best overall performance with SUnsET fine-tuned models within each dataset. The results on smaller models are more mixed across datasets, likely due to the difficulty for smaller models to learn the unstructured evidence task in general. Learning from SUnsET appears to be especially useful for improving summaries on multi-document datasets (SummHay and SQuALITY), which always see improvements over the unstructured baseline.

G Training Data Requirements

To observe the impact of number of SUnsET training samples on summary quality, we plot relevance and consistency vs. number of training samples for SQuALITY and ScholarQABench in Figure 23 and Figure 24. Interestingly, we find that performance generally peaks with only a modest amount of data (around 1k-3k samples depending on the model) at which point performance plateaus or slightly drops. It is likely that performance peaks when there is enough data to largely cover the distribution of data which is relevant for learning the task. Thus, more data does not result in more gains in performance, leading to the plateaus we see. We could potentially see additional performance gains by controlling the style of document generated, for example generating data which matches the target domain.

P2: Outline
Imagine that you must write a book. This book can be either fiction or non-fiction.

This is the title of your book: `{title}`

Please write an outline of this book. Please include the title of the book, and a list of chapters or sections that the book will contain. The book should have 6 sections or chapters.

OUTPUT FORMAT

Please output the outline as a JSON object where the keys are the chapters and the values are a brief outline of the chapter.

In other words, as:

```python
{ 'Chapter 1': 'Chapter 1 outline',
'Chapter 2': 'Chapter 2 outline',
...
'Chapter N': 'Chapter N outline'
}
```

Figure 9: Outline generation prompt. The `{title}` field is replaced with the title of one document.

H Data Availability Statement

We create SUnsET in this work, as well as the code to generate SUnsET, which we release freely to the public under the MIT license.[5] The data are generated as sets of fiction and non-fiction books in English.

I Model Descriptions

Table Table 8 presents the full set of Huggingface model identifiers for the LLMs used in our experiments. The model cards containing relevant information on number of parameters, context length, vocabulary size, etc. are available on their model page on the Huggingface website. All training and inference are performed using 1-2 Nvidia A100 GPUs with 48GB of memory. Prior to training we ran a brief hyperparameter search to find the parameters used in this study, sweeping over the following values (selected values in **bold**):

- Learning rate: [1e-6, 5e-4] (**5e-5**)
- Batch size: {**2**, 4, 8, 16, 32}
- Warmup steps: {0, **10**, 50, 100, 150, 200, 300}
- Train epochs: {1, 2, 3, 4, 5, 8, **10**, 12, 20}
- Lora rank: {2, 4, 8, 12, **16**, 32}

J Software Package Parameters

- NLTK (Bird, 2006): We use the punkt sentence tokenizer for sentence tokenization
- VLLM: We use top p sampling at 90% with a temperature of 1. for inference. We set maximum new generated tokens to 2,000
- OpenAI GPT 4o Mini: We use top p sampling at 90% with a temperature of 1 for all prompts except title generation (temperature set to 1.2) and filtering (deterministic highest probability token output).
- DeepSeek-V3: We use top p sampling at 90% with a temperature of 1 for all prompts.

K Evaluation Robustness

We use autoraters (i.e. LLM as a judge) for much of our evaluation. While we use a previously validated prompting and modeling setup (Liu et al., 2024b), we use DeepSeek-V3 as our autorater due to its high performance and low cost. We validated the robustness of DeepSeek-V3 as an autorater by taking a sample of 710 outputs summaries from

[5] https://anonymous.4open.science/r/sunset-BD72/README.md

P3.1: Queries Prompt
Imagine that you must write a book. You are given the following outline of the book

`{outline}`

Please write a list of 5 questions about the book which summarize the book.

Please try to cover different general aspects of the content.

Please make the questions very concise.

OUTPUT FORMAT

Please separate each question with a single newline character ("\n")

Figure 10: Query generation prompt. The `{outline}` is filled with the outline generated by Figure 9.

our evaluation and re-evaluating them with GPT 4o Mini (Liu et al., 2023). We measure the Pearson's R correlation between the ratings (2 ratings per summary) given by GPT 4o mini and DeepSeek-V3, finding a strong correlation of 73.29. This indicates the robustness of our evaluation which relies on DeepSeek-V3.

	SLT^{S}		LAS^{S}		SMH^{M}		SQB^{M}	
Model	Rel_{Prec}	Con_{Prec}	Rel_{Prec}	Con_{Prec}	Rel_{Prec}	Con_{Prec}	Rel_{Prec}	Con_{Prec}
Llama 3.2 1B	12.50	12.50	30.94	20.51	50.00	0.00	37.50	50.00
Fixed Gran.	19.86	4.10	39.22	25.86	25.94	8.88	21.82	11.47
+ SUnsET	18.80	10.61	41.27	32.05	0.00	0.00	45.18	24.08
+ Shuffled	28.60	13.01	50.34	48.86	50.00	0.00	62.38	48.20
Llama 3.2 3B	34.27	20.34	62.30	55.77	54.34	44.53	52.39	39.86
Fixed Gran.	34.84	15.24	62.02	56.35	24.59	24.91	35.86	29.97
+ SUnsET	45.17	25.65	61.16	53.96	64.75	59.25	52.91	45.00
+ Shuffled	44.28	27.20	62.76	54.42	65.76	62.84	60.98	56.37
Llama 3.1 8B	42.69	27.70	67.18	61.79	62.72	57.14	49.95	39.24
Fixed Gran.	44.45	26.84	59.66	54.80	39.14	39.00	50.21	49.70
+ SUnsET	50.91	33.71	75.21	70.45	**74.31**	70.96	67.36	61.17
+ Shuffled	53.13	36.79	**73.78**	68.99	70.55	67.15	64.70	61.12
Mistral Nemo 2407	31.67	14.00	60.27	53.41	**73.78**	**73.78**	69.49	61.38
Fixed Gran.	32.44	19.12	60.28	54.00	29.59	25.97	37.86	28.03
+ SUnsET	**57.34**	36.90	78.96	78.69	73.62	70.84	**71.44**	66.50
+ Shuffled	56.07	38.18	78.97	78.39	70.58	65.37	64.97	61.20
Mixtral 8x7B	47.82	32.79	81.58	83.76	68.54	66.53	53.67	48.02
Fixed Gran.	43.78	24.11	64.14	61.01	37.43	29.62	61.32	**67.63**
+ SUnsET	50.74	35.96	82.94	82.94	69.77	69.82	60.82	57.49
+ Shuffled	52.52	**38.71**	**84.19**	**85.29**	73.80	73.33	61.94	59.22
GPT 4o Mini	*60.11*	*52.11*	*77.92*	*74.76*	*77.09*	*75.57*	*57.49*	*49.18*

Table 4: Relevance and consistency **precision** of evidence sentences with respect to their citances. Precision measures the average citation quality within a given summary. **Bold** indicates best overall performance, Underline indicates best performance for individual models. ^{S} indicates single document tasks, ^{M} indicates multi-document. SQ is SQuALITY, LAS is LexAbSumm, SMH is SummHay, and SQB is ScholarQABench

P3.2: Initial Summaries and Evidence
Imagine that you are writing a book. This is an outline of the book

`{outline}`

Please address the following question about the book:

`{question}`

Please write a summary which addresses the question. Please make the summary as specific and detail oriented as possible. Please include actual examples from the book when possible. Please do not write more than is absolutely necessary.

After you write the summary, please write exact quotes and passages you will include in the book, from which the summary could be written. Please include at least `{n_evidence}` of these passages, which you intend to include verbatim in the book. Please indicate the exact chapter where the passages will be written in a separate field.

OUTPUT FORMAT

Please a JSON object with two fields: "summary", "evidence", and "chapter". The summary field should have the summary. The evidence field should have a list of evidence sentences from the book. The chapter field should have the exact chapter where the corresponding evidence sentence will appear. Please only indicate the chapter number for this field. There should be the same number of elements in the "evidence" field as there are in the "chapter" field. In other words, as:

```python
{
'summary': 'Summary text',
'evidence': ['evidence sentence 1', 'evidence sentence 2', ...]
'chapter': [1, 4, ...]
}
```

Figure 11: Initial summary and evidence generation prompt. The `{outline}` and `{question}` fields are filled by the output of the previous prompts, while the `{n_evidence}` field is filled by a random number between 5 and 10.

P4.1: Document Section Generation
Imagine that you must write a book. You are given the following outline of the book

```
{outline}
```

Please write the following chapter of the book in its entirety:

```
{chapter}
```

Please also include the following sentences somewhere in the chapter. You must include these passages verbatim (i.e., EXACTLY as is). It is imperative that you do this, otherwise the book will be incomplete:

```
{evidence}
```

OUTPUT FORMAT

Please wrap the content of the chapter you write in a markdown codeblock, in other words, like:

```
content
```

Figure 12: Document section generation prompt. The `{chapter}` field is filled by the title of the section being generated, as given in the outline.

P4.2: Evidence Retrieval Prompt
Please read the following book chapter:

```
{chapter}
```

The following passage should have been included in the chapter but was not:

```
{passage}
```

Please retrieve the passage from the chapter which is CLOSEST to the given passage.

OUTPUT FORMAT

Please wrap the passage in a markdown codeblock, in other words, like:

```
passage
```

Figure 13: Prompt to retrieve evidence from the document when previously generated evidence is not included verbatim. The `{passage}` field is filled with one piece of evidence that was supposed to be included in the section.

P5.1: Refinement Prompt
Imagine that you are giving an exam about a book. This is the book

```
{book}
```

On an exam, you are asked to summarize the book with respect to this question:

```
{question}
```

This is the summary that you are grading:

```
{summary}
```

Please rewrite this response so that it is totally accurate and fully addresses the question.

Please make the response as specific and detail oriented as possible. The following passages from the document should help in crafting the response:

```
{passages}
```

OUTPUT FORMAT

Please wrap the content of the summary you write in a markdown codeblock, in other words, like:

```
content
```

Figure 14: Summary refinement prompt after content has been generated. The `{book}` field is filled with the entire document, where each section is concatenated together. Other fields are filled with the output from the previous prompts.

P5.2: Citance generation
Imagine that you have written a research essay about a book. You have also extracted passages from the book which you used to write the essay.

Your job is to add citations to the essay which properly reference the passages that you have extracted.

Here is the essay:

`{essay}`

And here are the evidence passages from the book, each of which is given a number:

`{evidence}`

Please add citations to all citation-worthy statements in the essay using the numbered evidence list, by indicating the citation numbers of the corresponding evidence. More specifically, add the citation number at the end of each relevant sentence in the essay before the punctuation mark e.g., 'This work shows the effectiveness of problem X [1].' when the passage [1] in the evidence list provides full support for the statement. Only add a citation if it is fully relevant and unambiguously supportive of that sentence. Not all evidences may be relevant, so only cite those that directly support the statement. Please do not add any explanations or justifications for the evidence, simply indicate the evidence numbers if they are relevant. If a sentence does not use any of the provided evidence, please simply copy the sentence as is and do not add anything to the end of it. If multiple evidences support a statement, please cite them together (e.g., [1][2]). For each citation-worthy statement, you only need to add at least one citation, so if multiple evidences support the statement, just add the most relevant citation to the sentence.

Figure 15: Prompt to add citation references to sentences based on extracted evidence. The `{essay}` field is filled with a summary and the `{evidence}` field is filled with its corresponding evidence.

P6: Validation Prompt
Imagine that you are judging the quality of a summary of a book. This is the book

`{book}`

Here is a question about the book:

`{question}`

And here is the summary which addresses the question:

`{summary}`

Please judge if you think that the summary meets ALL of the following criteria:

1) The summary is absolutely faithful to the book (in other words, all of the information in the summary is contained in the book)

2) The summary FULLY addresses the question

Please think carefully about your answer. If you think that ALL of the criteria are met, please simply respond with "YES".

Otherwise, please simply respond with "NO".

Figure 16: Prompt to add citation references to sentences based on extracted evidence. Fields are filled with the output of previous prompts.

Baseline Non-Pipelined Prompt
Imagine that you must write a book. This book can be either fiction or non-fiction.

You can select any subject to write your book about. Please make the book interesting.

Please perform the following tasks and output everything in as a JSON object:

Please write the title of the book.
`{title_prompt}`

Then, please write an outline of this book. Please include a list of chapters or sections that the book will contain. The book should have 6 sections or chapters.

Then, please write a list of 5 questions about the book which summarize the book.

Then, please write a summary for each question which addresses the question.

Then, please write the entire contents of the book. The book should be long, and you should write out the ENTIRE content.

Then, extract specific passages from the book for each summary which serve as evidence for the summary.

OUTPUT FORMAT
Please create a well-formatted JSON object with the following fields:

title: The title of the book (formatted as a string)
outline: The outline of the book (formatted as a string)
questions: The questions about the book (formated as a list)
summaries: The summaries addressing each question (formatted as a list of the same length as "questions")
document: The full book (formatted as a string)
evidence: A list of evidence passages (formatted as a list of the same length as "questions")

Figure 17: Baseline non-pipelined prompt that we use as a point of comparison. The field `{title_prompt}` is empty for the baseline without diversity enforced, and filled with a list of previous titles and the prompt "Please do not use any of the following titles:".

Training and Inference Prompt

Your task is to read a document and then write an essay which addresses the following question: `{question_text}`

To write your essay, you should read the document and identify key passages which will help guide your response. Extract every passage which is directly relevant for your essay. Please copy each extracted passage to a list in the format specified below. Please copy the exact text of each passage (do NOT paraphrase!). Then, write your essay which addresses the query.

Please add citations to all citation-worthy statements using the extracted evidence, by indicating the citation numbers of the corresponding evidence. More specifically, add the citation number at the end of each relevant sentence before the punctuation mark e.g., 'This work shows the effectiveness of problem X [1].' when the passage [1] in the evidence list provides full support for the statement. Only add a citation if it is fully relevant and unambiguously supportive of that sentence. Not all evidences may be relevant, so only cite those that directly support the statement. Please do not add any explanations or justifications for the evidence, simply indicate the evidence numbers if they are relevant. If a sentence does not use any of the provided evidence, please simply copy the sentence as is and do not add anything to the end of it. If multiple evidences support a statement, please cite them together (e.g., [1][2]). For each citation-worthy statement, you only need to add at least one citation, so if multiple evidences support the statement, just add the most relevant citation to the sentence.

Please limit to only 10 pieces of evidence.

Here is the document: `{context}`

OUTPUT FORMAT
Output your response as:
EVIDENCE:
[1] Extracted passage 1
[2] Extracted passage 2
...
[N] Extracted passage N
RESPONSE:
response

Figure 18: Full prompt used for fine-tuning and inference. The `{question_text}` field is filled with a single query, and the `{context}` field is filled with the document context.

Summary Combination Prompt

Here is a list of summaries of different sections of a document with respect to the query "{question_text}":

{context}

Please combine these summaries into a single summary which addresses the query. If a summary mentions that the query is not addressed, please ignore that summary. Please keep all relevant citations in the final summary. Here is a list of the original citations:

{evidence}

Figure 19: Prompt to combine section summaries into one final summary.

	SLT^{S}		LAS^{S}		SMH^{M}		SQB^{M}	
Model	Rel_{Rec}	Con_{Rec}	Rel_{Rec}	Con_{Rec}	Rel_{Rec}	Con_{Rec}	Rel_{Rec}	Con_{Rec}
Llama 3.2 1B	0.10	0.10	0.94	0.69	0.27	0.00	0.06	0.08
Fixed Gran.	0.33	0.12	5.24	3.42	0.28	0.15	1.88	1.14
+ SUnsET	0.82	0.43	4.06	2.45	0.00	0.00	2.40	0.93
+ Shuffled	1.26	0.52	2.01	1.94	0.05	0.00	0.48	0.41
Llama 3.2 3B	4.85	2.82	11.64	10.13	5.75	4.90	11.22	8.36
Fixed Gran.	18.13	7.45	39.63	35.85	0.93	0.78	24.02	20.37
+ SUnsET	20.14	11.86	26.95	23.70	26.68	24.54	10.18	8.80
+ Shuffled	11.09	6.85	14.56	12.55	22.24	20.82	11.53	11.07
Llama 3.1 8B	8.90	5.61	22.41	20.76	25.52	23.23	16.68	13.17
Fixed Gran.	14.88	8.98	36.83	33.73	12.22	12.19	33.55	32.60
+ SUnsET	21.32	14.28	**41.31**	**38.72**	**47.39**	**45.45**	**35.28**	32.47
+ Shuffled	16.80	11.70	35.13	32.78	42.35	40.44	32.31	30.86
Mistral Nemo 2407	0.47	0.20	1.13	1.08	5.18	5.17	4.94	4.54
Fixed Gran.	5.39	3.26	10.40	9.34	2.64	2.39	12.04	8.79
+ SUnsET	17.48	11.30	19.93	19.66	16.63	15.80	17.68	16.59
+ Shuffled	13.81	9.38	19.59	19.14	16.17	15.06	13.54	13.00
Mixtral 8x7B	15.47	11.04	29.99	30.85	29.87	28.54	13.92	12.46
Fixed Gran.	**33.32**	**18.68**	36.40	34.42	6.32	5.75	34.11	**37.82**
+ SUnsET	19.06	13.64	30.65	30.68	37.91	37.31	23.06	21.80
+ Shuffled	20.40	15.40	31.82	32.08	39.55	38.65	27.00	26.22
GPT 4o Mini	*28.38*	*23.86*	*51.15*	*49.07*	*55.03*	*53.93*	*25.82*	*21.99*

Table 5: Relevance and consistency **recall** of evidence sentences with respect to their citances. Recall measures citation quality and averages based on the total number of sentences in a summary. This penalizes models which produce fewer citations. **Bold** indicates best overall performance, Underline indicates best performance for individual models. ^{S} indicates single document tasks, ^{M} indicates multi-document. SQ is SQuALITY, LAS is LexAbSumm, SMH is SummHay, and SQB is ScholarQABench

Relevance Prompt
You will be given one summary written for a document based on a query about that document.

Your task is to rate the summary on one metric with respect to the query.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: Relevance (1-5) - selection of important content from the source. The summary should include only important information from the source document which is relevant for the query. Annotators were instructed to penalize summaries which contained redundancies, excess information, and information which does not address the query.

Evaluation Steps:

1. Read the query, the summary, and the source document carefully.
2. Compare the summary to the query and the source document and identify the main point of the document which is relevant to the query.
3. Assess how well the summary covers the main points of the source document which are relevant to the query, and how much irrelevant or redundant information it contains.
4. Assign a relevance score from 1 to 5.

Example:

Source Text:

`{document}`

Query:

`{query}`

Summary:

`{summary}`

Evaluation Form (scores ONLY): - `{Relevance}`

Figure 20: Relevance evaluation prompt from (Liu et al., 2024b). The `{document}` field is filled with the document context and the `{summary}` field is filled with a summary. When used to evaluate summarization, the `{query}` field is filled with the query used to generate the summary. For citation evaluation, the `{query}` field and all references to queries are removed from the prompt.

Consistency Prompt
You will be given one summary written for a document based on a query about that document.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Consistency (1-5) - the factual alignment between the summary and the summarized source with respect to the query. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.

Evaluation Steps:

1. Read the source document carefully and identify the main facts and details it presents with respect to the query.
2. Read the summary and compare it to the source document. Check if the summary contains any factual errors that are not supported by the source document.
3. Assign a score for consistency based on the Evaluation Criteria.

Example:

Source Text:

`{document}`

Query:

`{query}`

Summary:

`{summary}`

Evaluation Form (scores ONLY): - `{Consistency}`

Figure 21: Consistency evaluation prompt from (Liu et al., 2024b). The `{document}` field is filled with the document context and the `{summary}` field is filled with a summary. When used to evaluate summarization, the `{query}` field is filled with the query used to generate the summary. For citation evaluation, the `{query}` field and all references to queries are removed from the prompt.

	SLT^{S}		LAS^{S}		SMH^{M}		SQB^{M}	
Model	Rel_{F1}	Con_{F1}	Rel_{F1}	Con_{F1}	Rel_{F1}	Con_{F1}	Rel_{F1}	Con_{F1}
Llama 3.2 1B	0.14	0.14	1.22	0.84	0.36	0.00	0.11	0.14
Fixed Gran.	0.40	0.13	6.67	4.40	0.39	0.19	2.27	1.35
+ SUnsET	1.18	0.62	5.43	3.59	0.00	0.00	3.13	1.26
+ Shuffled	1.85	0.80	3.14	3.04	0.08	0.00	0.85	0.72
Llama 3.2 3B	6.61	3.86	15.17	13.29	7.66	6.52	14.12	10.54
Fixed Gran.	21.71	9.02	45.80	41.44	1.37	1.13	27.77	23.49
+ SUnsET	25.36	14.76	33.42	29.40	32.21	29.59	13.76	11.85
+ Shuffled	15.14	9.33	19.45	16.80	26.78	25.15	17.45	16.55
Llama 3.1 8B	11.66	7.38	28.89	26.76	32.07	29.17	20.73	16.32
Fixed Gran.	18.90	11.32	42.44	38.86	14.29	14.23	38.56	37.64
+ SUnsET	27.69	18.48	**50.78**	**47.62**	**53.62**	**51.43**	**44.03**	40.49
+ Shuffled	23.13	16.12	44.16	41.18	48.72	46.50	41.49	39.59
Mistral Nemo 2407	0.53	0.23	1.36	1.29	6.68	6.68	6.08	5.54
Fixed Gran.	6.61	3.93	13.36	11.95	3.71	3.36	15.05	11.03
+ SUnsET	21.71	13.99	23.38	23.09	20.73	19.71	22.00	20.61
+ Shuffled	17.67	11.96	22.85	22.42	19.82	18.38	16.87	16.14
Mixtral 8x7B	17.83	12.64	34.27	35.23	33.40	32.02	17.30	15.48
Fixed Gran.	**36.35**	**20.33**	42.34	40.15	8.45	7.46	40.06	**44.40**
+ SUnsET	22.60	16.11	35.81	35.81	42.91	42.27	28.61	26.94
+ Shuffled	23.79	17.85	37.21	37.57	43.89	42.98	32.25	31.16
GPT 4o Mini	*37.39*	*31.70*	*61.17*	*58.68*	*63.61*	*62.35*	*33.71*	*28.63*

Table 6: Relevance and consistency **F1** of evidence sentences with respect to their citances. We follow a similar setup to (Laban et al., 2024; Asai et al., 2024) where we measure citation precision and recall in order to calculate an overall F1 score for both relevance and consistency. **Bold** indicates best overall performance, Underline indicates best performance for individual models. ^{S} indicates single document tasks, ^{M} indicates multi-document. SQ is SQuALITY, LAS is LexAbSumm, SMH is SummHay, and SQB is ScholarQABench

	SLT^{S}		LAS^{S}		SMH^{M}		SQB^{M}	
Model	Rel	Con	Rel	Con	Rel	Con	Rel	Con
Llama 3.2 1B	2.28	1.63	3.09	2.88	3.52	3.70	2.90	2.93
Fixed Gran.	2.42	1.49	3.28	2.81	3.09	3.32	3.28	3.36
+ SUnsET	2.60	2.23	2.99	2.75	3.82	4.04	3.17	3.02
+ Shuffled	2.57	2.15	3.06	2.78	3.83	4.35	3.18	3.07
Llama 3.2 3B	3.66	3.52	4.26	4.49	4.47	4.83	3.99	4.21
Fixed Gran.	3.40	3.11	4.12	4.34	3.45	3.53	4.04	4.28
+ SUnsET	3.49	3.10	4.13	4.17	4.73	4.91	4.26	4.20
+ Shuffled	3.16	2.68	4.17	4.13	4.88	4.95	4.36	4.20
Llama 3.1 8B	4.26	4.44	4.60	**4.81**	4.84	4.92	4.07	4.24
Fixed Gran.	4.23	4.34	4.59	4.79	4.43	4.55	4.52	4.59
+ SUnsET	4.23	4.24	4.65	**4.81**	4.89	**4.98**	4.58	4.55
+ Shuffled	4.08	4.02	**4.66**	4.75	**4.92**	**4.98**	**4.68**	**4.69**
Mistral Nemo 2407	4.15	4.15	3.52	3.70	4.05	4.37	3.09	3.25
Fixed Gran.	4.12	4.26	4.42	4.68	2.54	2.62	4.06	4.23
+ SUnsET	4.29	4.31	4.24	4.39	4.52	4.66	3.65	3.77
+ Shuffled	4.41	4.38	4.35	4.46	4.50	4.73	3.76	3.86
Mixtral 8x7B	4.21	4.47	4.43	4.73	4.46	4.67	4.09	4.27
Fixed Gran.	4.46	4.63	4.46	4.71	3.93	4.08	4.19	4.43
+ SUnsET	4.48	4.64	4.54	4.79	4.49	4.74	4.29	4.43
+ Shuffled	**4.55**	**4.67**	4.56	**4.81**	4.55	4.78	4.20	4.43
GPT 4o Mini	*4.77*	*4.85*	*4.87*	*4.93*	*4.98*	*5.00*	*4.93*	*4.94*

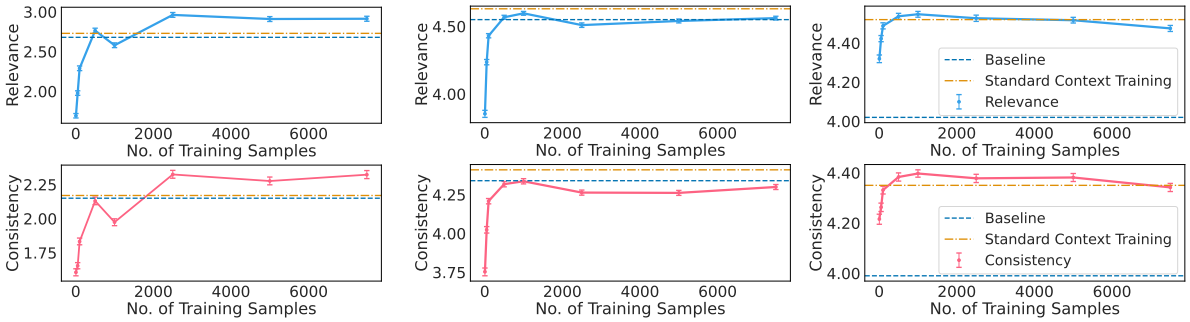
Table 7: Relevance and consistency of generated summaries. Relevance and consistency are measured using an autorater (DeepSeek-V3) (Liu et al., 2023) based on previously validated prompts (Liu et al., 2024b). **Bold** indicates best overall performance, Underline indicates best performance for individual models. ^{S} indicates single document tasks, ^{M} indicates multi-document. SQ is SQuALITY, LAS is LexAbSumm, SMH is SummHay, and SQB is ScholarQABench.

Example Document Snippet
Title: "Writing the Unwritable"
...They demonstrate that while writing the unwritable is fraught with difficulty, it can also yield transformative insights that resonate profoundly with readers. Writing the unwritable requires a recognition of the limitations of language, and a willingness to push against those boundaries. This requires not merely acceptance of silence or ambiguity but a bold declaration that some truths demand to be told, no matter how fraught the endeavor may be....
Example Query
What does it mean to write the unwritable, and what historical examples illustrate this concept?
Example Summary Snippet
To write the unwritable involves confronting and articulating subjects and experiences that resist verbal expression, often due to limitations of language, social taboos, and the impact of censorship [1][2][3].
Example Evidence Snippet
[1] Writing the unwritable requires a recognition of the limitations of language, and a willingness to push against those boundaries.

Figure 22: Snippets from a SUnsET document.

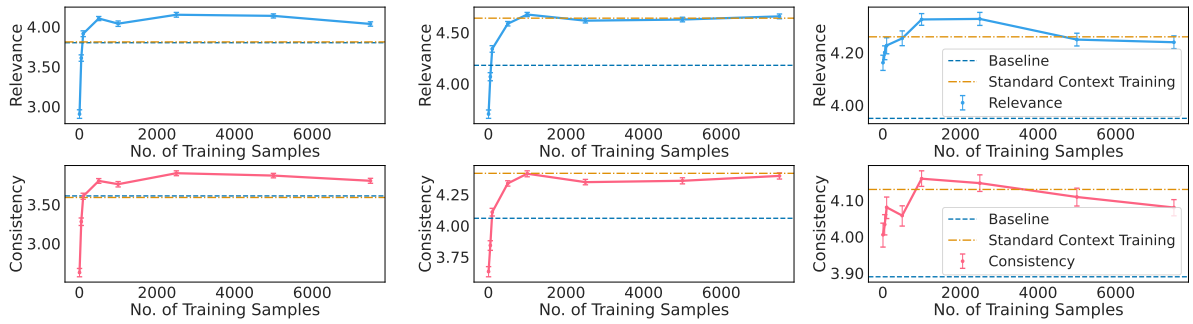
Model	Huggingface Identifier
Llama 3.2 1B	`meta-llama/Llama-3.2-1B-Instruct`
Llama 3.2 3B	`meta-llama/Llama-3.2-3B-Instruct`
Llama 3.1 8B	`meta-llama/Meta-Llama-3.1-8B-Instruct`
Mistral Nemo 2407	`mistralai/Mistral-Nemo-Instruct-2407`
Mixtral 8x7B	`mistralai/Mixtral-8x7B-Instruct-v0.1`

Table 8: Huggingface identifiers for models used in our experiments.



(a) Llama 3.2 1B (b) Llama 3.1 8B (c) Mixtral 8x7B

Figure 23: SQuALITY: Relevance and consistency performance vs. number of synthetic training samples.



(a) Llama 3.2 1B (b) Llama 3.1 8B (c) Mixtral 8x7B

Figure 24: ScholarQABench: Relevance and consistency performance vs. number of synthetic training samples.