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. Extracting and properly citing evidence spans could help improve the transparency and reliability of these summaries. At the same time, LLMs suffer from positional biases in terms of which information they understand and attend to, which could affect evidence citation. Whereas previous work has focused on evidence citation with predefined levels of granularity (e.g. sentence, paragraph, document, etc.), we propose the task of long-context query focused summarization with *unstructured* evidence citation. We show how existing systems struggle to generate and properly cite unstructured evidence from their context, and that evidence tends to be "lost-in-the-middle". To help mitigate this, we create the **S**ummaries with **Uns**tructured **E**vidence **T**ext dataset (SUnsET), a synthetic dataset generated using a novel domain-agnostic pipeline which can be used as supervision to adapt LLMs to this task. We demonstrate across 5 LLMs of different sizes and 4 datasets with varying document types and lengths that LLMs adapted with SUnsET data generate more relevant and factually consistent evidence than their base models, extract evidence from more diverse locations in their context, and can generate more relevant and consistent summaries.

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, sets of research papers, and long legal documents, and summarize them based on user queries (Koh et al., 2023; Su et al., 2024; Beltagy et al., 2020; Reid et al., 2024). The difficulty of this task lies in the need to attend to relevant information in the source document(s) given a query and simultaneously derive coherent, factually consistent, and distilled insights (Goldman et al., 2024). While LLMs have achieved much progress 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). While progress has been made on structured evidence citation (i.e., with fixed levels of granularity such as sentences and documents) (Li et al., 2023), in order to improve the flexibility and explainability of long-context query focused summaries we propose to study the task of *unstructured* evidence citation.

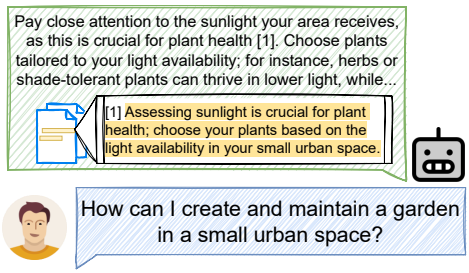


Figure 1: We study query focused summarization with *unstructured* evidence citation from long contexts

Within this, we explore two key barriers to achieving good performance with reliable summaries. First, LLMs have positional biases (Liu et al., 2024c; Ravaut et al., 2024), focusing on the earlier and later tokens in their input context (Zhang et al., 2024b). This can potentially bias which evidence a model selects to support a summary. Second, fine-tuning has benefits in terms of inference efficiency (Wu et al., 2024), and in some cases can outperform inference interventions (Zhang et al., 2024a; Liu et al., 2022), but would require a large dataset with specialized examples of long documents, queries, extracted evidence, and summaries which cite this evidence. Creating such a dataset requires an extensive amount of time, money, and expertise (Asai et al., 2024; Laban et al., 2024; Santosh et al., 2024).

To address these issues, we provide the first study on unstructured (i.e., no fixed level of granularity) evidence citation in long-context query-focused summarization. While attribution has been investigated in recent works (Laban et al., 2024; Asai et al., 2024; Li et al., 2023), we present a novel and more challenging scenario where a model must extract unstructured text spans from its context to use as supporting information for its summary. We show for base models that extracting and citing unstructured evidence is challenging, and that evidence is often lost-in-the-middle. To help alleviate this, we present a fine-tuning approach based on an inductively generated synthetic dataset called the **S**ummaries with **Uns**tructured **E**vidence **T**ext dataset (SUnsET). Across 5 different models and 4 different datasets (single- and multi-document), we demonstrate that fine-tuning on SUnsET data can help mitigate lost-in-the-middle, increase citation accuracy and coverage, and improve summary quality. We release SUnsET and our generation code to the public for further study.[1]

2 Challenges in Query Focused Long Context Summarization

Query focused, long context summarization 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). The main focus of our work is on evidence attribution (Laban et al., 2024; Worledge et al., 2024; Li et al., 2023) and its relation to the lost-in-the-middle problem (Liu et al., 2024c; Ravaut et al., 2024).

2.1 Evidence Attribution

Though LLMs can generate convincing summaries, in practice people often prefer to acquire information through media where provenance is fully transparent (for example, a normal search engine) (Worledge et al., 2024). 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, including SummHay (Laban et al., 2024) and OpenScholar (Asai et al., 2024). However, most works focus on structured attribution with fixed levels of granularity (e.g., sentences, paragraphs, or documents) (Li et al., 2023). To the best of our knowledge, we provide a first study on unstructured evidence citation in long context query focused summarization, which is more flexible than the fixed-granularity approach.

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 further 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 Cite and Summarize

Our task is: given a query about a long input consisting of one or more documents, generate a response to the query which cites *unstructured* evidence from the input. We differentiate ourselves from previous work on summarization with attribution (Laban et al., 2024; Asai et al., 2024; Li et al., 2023) by requiring a model to extract unstructured text as evidence with no predefined levels of granularity. This evidence must also be relevant and consistent with both the document and the summary sentences. While more challenging, this en-

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

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. Indicate which sections the passages should be included in.
P4. Document sections: Generate each section of the document one at a time. Ensure that evidence passages are included verbatim in each section.
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 and is faithful to the document, and includes inline attribution to evidence passages.

Figure 2: Six stage inductive data generation pipeline. The full prompts for each stage are given in Appendix A Figure 8 - Figure 16.

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 3: Snippets from a SUnsET document.

ables flexible and explainable summary generation, and citations can easily be generated in both multi-document and single-document settings.

To tackle this we create SUnsET, a synthetic dataset which is generalizable across domains, using a novel inductive generation pipeline. We then use this data to fine-tune models using adapters (Houlsby et al., 2019) to improve unstructured evidence citation and summary quality, as well as 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 which use low-rank adapters (LoRA) (Hu et al., 2022) trained on different views of SUnsET.

3.1 Generating SUnsET

Recent works have demonstrated the promise of generating large scale synthetic datasets for fine-tuning task specific models (Ziegler et al., 2024; Honovich et al., 2023; Wang et al., 2023; Chen et al., 2024; Xu et al., 2024) with the help of strong LLMs such as GPT-4 (OpenAI, 2023). Inspired by this, we propose SUnsET, a generic, domain-agnostic dataset generated from a novel inductive synthetic data pipeline that allows us to fine-tune downstream models to generate relevant and consistent query focused summaries from long-contexts with unstructured evidence citations.

Our pipeline generates datasets composed of long documents paired with queries and long-form answers to those queries. Each summary includes inline citations that reference relevant unstructured 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.

	SUnsET				Base				Base w/ Titles			
Method	Titles	Questions	Summaries	Documents	Titles	Questions	Summaries	Documents	Titles	Questions	Summaries	Documents
Moving Avg. TTR	0.816	0.751	0.836	0.820	0.387	0.670	0.797	0.350	0.588	0.631	0.778	0.352
Avg. Cosinse Dist.	0.780	0.806	0.733	0.682	0.425	0.725	0.716	0.042	0.607	0.660	0.610	0.040
Avg. Length (in words)	5.44	13.45	226.5	3767.4	6.65	9.85	23.79	474.8	5.76	10.21	24.45	433.8

Table 1: Statistics and diversity metrics of synthetically generated data.

The full generation process is described in Figure 2, 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 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 evidence positional biases. Following this, the queries, summaries, and evidence are refined in order to fully reflect the final document. Finally, we filter the summaries and evidence by prompting GPT 4o mini to predict if the summaries fully address the query and are fully supported by the document (see Figure 3 for an example).

To validate the pipeline, we additionally generate two baseline datasets. The first is generated by combining all the steps listed in Figure 2 into a single prompt. The second includes a control where we enforce that no repeated titles are generated (see Figure 17 in Appendix A). We compare each dataset using samples of 100 documents along dataset diversity metrics (average type-token ratio (TTR)(Bestgen, 2023), embedding cosine distance, and average word length) in Table 1. Baseline non-pipelined approaches produce shorter documents and shorter summaries, and these documents and summaries tend to be much more similar to each other than those generated using our pipeline.

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). One of the benefits of SUnsET documents is that they are highly modular, as the generated documents have both 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 the 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).

3.3 Summarizing with Unstructured Evidence

To generate summaries with unstructured citations, we design a prompt that is constructed of elements from previous work (Asai et al., 2024). The full prompt is given in Figure 18 in Appendix A. This prompt was refined through several rounds of experimentation, while the aspects related to citation formatting are taken from (Asai et al., 2024). We use this prompt both for inference and for supervised fine-tuning on SUnsET. After generating responses using this prompt, we validate that the format instructions are followed in order to separate evidence from the response. When the output is misformatted, we regenerate samples either until the format is correct, or until 5 attempts are reached. If the formatting is still in error, then we use the output of the last attempt as is for the summary. To deal with long contexts, we take a divide-and-conquer approach, which chunks each document according to the model's maximum content length, summarizes each chunk, and finally summarizes each summary. Thus, the output for each ⟨document, query⟩ pair is a ⟨summary, evidence_list⟩ pair containing the summary and a list of unstructured text spans from the context.

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

Table 2: Hallucination rates for evidence extraction. We directly 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 in the context.

	SLT^S		LAS^S		SMH^M		SQB^M	
Method	Rel_{F1}	Con_{F1}	Rel_{F1}	Con_{F1}	Rel_{F1}	Con_{F1}	Rel_{F1}	Con_{F1}
Llama 3.2 1B	0.00	0.00	0.94	1.06	**0.00**	**0.00**	0.23	0.18
+ Standard	**0.63**	**0.53**	**4.80**	**4.56**	**0.00**	**0.00**	**1.84**	**1.68**
+ Shuffled	0.48	0.26	2.83	2.84	**0.00**	**0.00**	0.00	0.00
Llama 3.2 3B	11.21	10.16	15.08	14.64	8.64	8.75	12.37	12.99
+ Standard	**36.19**	**25.12**	**43.98**	**40.64**	**37.73**	**39.03**	**37.16**	**34.39**
+ Shuffled	23.38	15.33	36.19	31.26	32.73	33.46	31.36	26.73
Llama 3.1 8B	17.21	15.15	31.17	30.65	34.18	37.96	32.08	32.85
+ Standard	**35.21**	**25.34**	**52.64**	**47.79**	**56.82**	**57.50**	**45.26**	**41.13**
+ Shuffled	29.36	20.65	49.90	44.19	54.79	54.27	39.53	36.17
Mistral Nemo 2407	2.75	2.37	5.34	4.58	10.37	10.25	5.67	5.36
+ Standard	**34.24**	**24.45**	38.21	36.88	**23.54**	**25.13**	**7.15**	**7.56**
+ Shuffled	32.52	22.84	**39.94**	**38.57**	21.58	23.23	4.65	4.08
Mixtral 8x7B	24.45	19.15	39.48	40.08	44.01	43.44	25.97	25.61
+ Standard	30.54	25.11	38.27	38.08	**48.71**	**51.85**	38.37	38.59
+ Shuffled	**32.87**	**25.86**	**44.13**	**44.48**	46.67	49.09	**39.65**	**41.89**
GPT 4o Mini	*42.62*	*36.23*	*59.48*	*53.96*	*64.99*	*60.14*	*37.65*	*33.11*

Table 3: Relevance and consistency of evidence sentences with respect to their citances. Relevance and consistency are measured using an autorater (GPT-4o-mini) (Liu et al., 2023) based on previously validated prompts (Liu et al., 2024b). 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. ^{S} indicates single document tasks, ^{M} indicates multi-document. SQ is SQuALITY, LAS is LexAbSumm, SMH is SummHay, and SQB is ScholarQABench

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 B). 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).

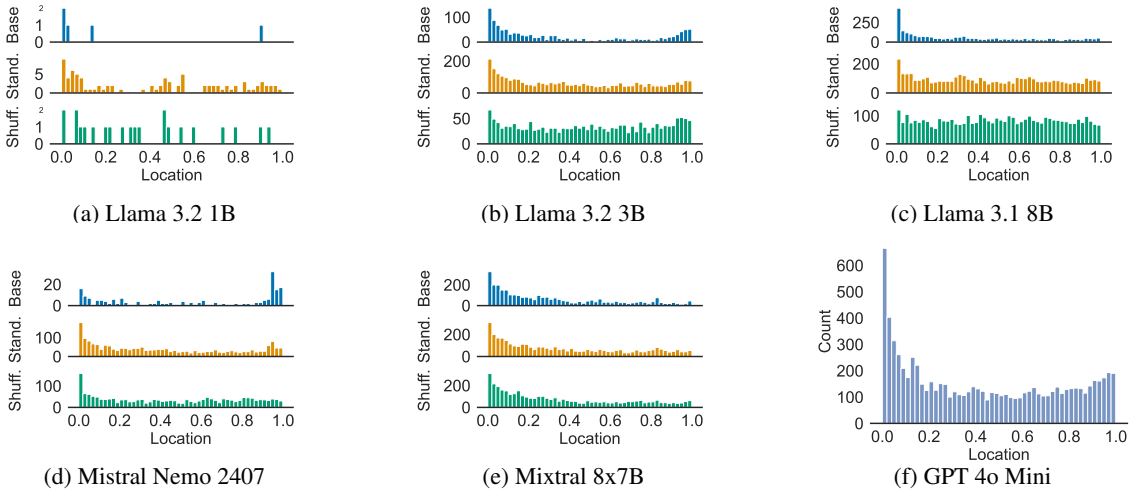
Models We use a set of LLMs covering multiple sizes and pretraining configurations. This includes Llama 3.2 1B, Llama 3.2 3B, Llama 3.1 8B (Dubey et al., 2024), Mistral Nemo 2407, and Mixtral 8x7B.[2] Additionally, we provide an upper bound estimate on performance using GPT 4o mini with no fine-tuning.

Evaluation For 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 using previously validated prompts listed in Appendix A (Figure 20 and Figure 21) (Liu et al., 2024b). 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 GPT-4o-mini (OpenAI, 2023).[3]

Training, and Inference We generate a total of 2,352 synthetic documents, giving us 11,309 ⟨document, question, summary⟩ tuples. We hold

[2]Huggingface model IDs are listed in Appendix D Table 5

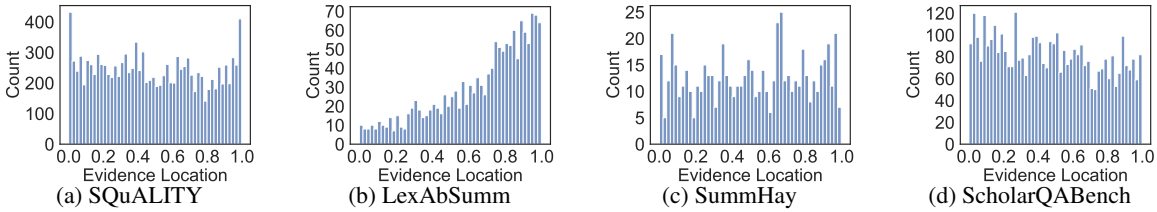
[3]We test the robustness of our evaluation in Appendix F



(a) Llama 3.2 1B (b) Llama 3.2 3B (c) Llama 3.1 8B

(d) Mistral Nemo 2407 (e) Mixtral 8x7B (f) GPT 4o Mini

Figure 4: Location of extracted evidence in the provided source context for different methods.



(a) SQuALITY (b) LexAbSumm (c) SummHay (d) ScholarQABench

Figure 5: Location of ground truth evidence in each dataset.

out 200 documents for validation and early stopping. In all cases we fine tune using the Huggingface Transformers implementation of LoRA (Hu et al., 2022) with a rank and α of 16 applied to all linear operators of each model.

4.1 RQ1: Can LLMs Generate Unstructured Evidence?

Using the datasets and models just described, we first test if LLMs can extract and effectively utilize unstructured evidence, as well as the impact of training on SUnsET. We look at two aspects of citation ability: evidence hallucination and evidence accuracy.

To study evidence hallucination, we attempt to match each sentence generated in the evidence list to its position in the context. We do so with two measures: exact string match and percent overlap of longest common substring (LCS) between the evidence and context. We present the rate of exact evidence match and 50% LCS overlap for all models aggregated across all datasets in Table 2. We see that **all base models struggle to faithfully copy evidence from the context**. This includes GPT 4o mini, which only faithfully copies 11% of the time. This rate is dramatically improved in all cases except for the smallest model (Llama 3.2 1B) by learning to cite unstructured evidence using SUnsET. Additionally, we see that rates of citation also dramatically increase (6.8$\times$ for Mixtral 8x7B). We find that learning to cite using SUnsET greatly improves the extraction of unstructured evidence from arbitrary contexts.

Next, we study attribution quality using a measure similar to the citation accuracy presented in Asai et al. (2024) but based on the relevance and consistency of evidence with their citing sentences (i.e. citances). To measure attribution quality, we propose two measures: Relevance F1 and Consistency F1. These are calculated as follows: for a given $\langle$summary, evidence_list$\rangle$ pair, we first sentence tokenize the summary and extract all citations from each citance. Then, we pair each citance with each piece of evidence that it cites and measure either the relevance or consistency of the evidence with respect to the citance. We normalize these scores (originally between 1 and 5) to the range $[0, 1]$. To obtain a measure of precision, we average these scores over all the citances in the summary. For recall, we average the scores over *every* sen-

	SLT^{S}		LAS^{S}		SMH^{M}		SQB^{M}	
Method	Rel	Con	Rel	Con	Rel	Con	Rel	Con
Llama 3.2 1B	2.68	2.15	**3.68**	**3.38**	4.53	4.40	3.80	3.61
+ Standard	2.73$^{=}$	**2.17**$^{=}$	3.25	2.93	4.53$^{=}$	4.44$^{=}$	3.81$^{=}$	3.59$^{=}$
+ Shuffled	**2.79***	2.15$^{=}$	3.41	3.03	**4.66***	**4.55***	**3.97***	**3.69***
Llama 3.2 3B	**4.39**	**4.05**	**4.40**	**4.19**	4.82	4.74	4.28	4.11
+ Standard	4.22	3.80	4.19	4.02	**4.90***	**4.85***	4.41*	4.21*
+ Shuffled	3.84	3.38	4.25	4.02	4.89*	4.84*	**4.49***	**4.23***
Llama 3.1 8B	4.55	4.34	**4.64**	**4.52**	4.88	4.78	4.18	4.06
+ Standard	**4.63***	**4.41***	4.53	4.44	4.94*	**4.93***	4.64*	**4.42***
+ Shuffled	4.59*	4.34$^{=}$	4.55	4.44	**4.97***	4.92*	**4.68***	4.41*
Mistral Nemo 2407	4.27	4.09	3.83	3.85	4.27	4.15	3.15	3.23
+ Standard	4.43*	4.24*	4.03*	4.04*	4.54*	4.47*	**3.79***	**3.75***
+ Shuffled	**4.53***	**4.35***	**4.18***	**4.12***	**4.65***	**4.49***	3.49*	3.41*
Mixtral 8x7B	4.02	3.99	4.28	4.22	4.78	4.68	3.95	3.89
+ Standard	**4.52***	4.35*	**4.45***	**4.40***	**4.84***	**4.72***	4.26*	4.13*
+ Shuffled	4.51*	**4.40***	4.44*	4.38*	4.79$^{=}$	4.68$^{=}$	**4.33***	**4.18***
GPT 4o Mini	*4.98*	*4.92*	*4.93*	*4.77*	*4.99*	*4.98*	*4.94*	*4.76*

Table 4: Relevance and consistency of generated summaries. Relevance and consistency are measured using an autorater (GPT-4o-mini) (Liu et al., 2023) based on previously validated prompts (Liu et al., 2024b). * indicates significance as measured by non-overlapping bootstrapped confidence intervals with the baseline. $^{=}$ indicates no significant difference from baseline. ^{S} indicates single document tasks, ^{M} indicates multi-document. SLT is SQuALITY, LAS is LexAbSumm, SH is SummHay, and SQB is ScholarQABench.

tence in the summary, thus penalizing summaries which do not use citations. F1 is then calculated as $\frac{2*p*r}{p+r}$.

We present results on citation relevance and consistency in Table 3. We again find that without any intervention, base models are generally bad at selecting and generating relevant and consistent evidence. As expected, larger models are better at this task, with base GPT 4o mini providing a modest upper bound on performance. We see that training on SUnsET helps to significantly close this gap and greatly improve citation quality. Smaller models see fewer gains, while larger models are able to adapt using SUnsET much more strongly, in some cases surpassing GPT 4o mini (e.g., Llama 3.1 8B for ScholarQABench). As with evidence hallucination, standard fine-tuning generally performs better, except for Mixtral, which sees a boost from doing shuffled training. Overall, we find that **base models struggle to utilize unstructured evidence, while SUnsET helps models to learn this across highly diverse test sets**.

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

Next, we explore to what extent this evidence is lost in the middle. To characterize positional bias, we match extracted evidence to its relative location in the document context (based on 50% LCS overlap) and plot this as a histogram in Figure 4. 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 in Figure 5.[4]

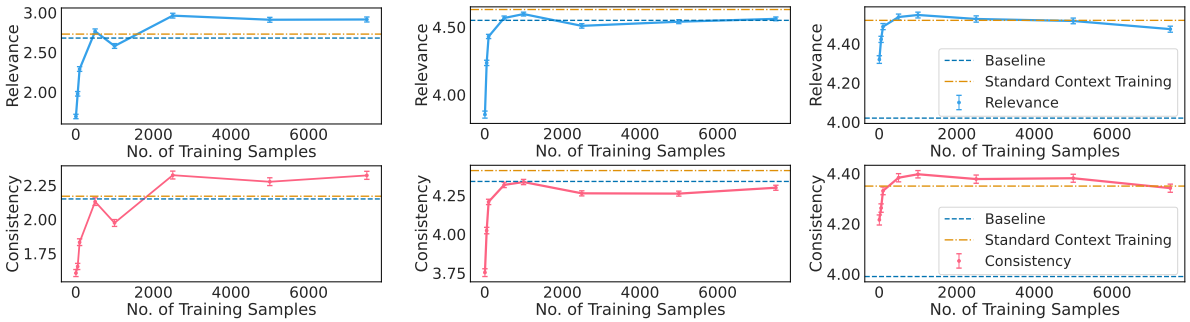
We find that **evidence is lost in the middle for all base models**. This includes GPT 4o Mini, which has a sharp spike of evidence in the early context. This stands in contrast to ground truth summary location distributions, which are uniform in all cases except for LexAbSumm which has a bias for evidence at the end of the context. In general, training on SUnsET without shuffling increases the rate of evidence extraction, but does not decrease the bias significantly. Shuffling on the other hand, increases the rate of evidence extraction and can decrease the bias. This is especially the case for Llama 3.1 8B and Llama 3.2 3B. Thus, similar to previous work on RAG for question-answering tasks (Zhang et al., 2024b), we find that shuffled training has the potential to help reduce positional biases for evidence extraction as well. This presents a tradeoff between training with shuffled and unshuffled documents: on the one hand, standard training leads to generally higher intrinsic citation and evidence quality, while on the other, fails to reduce positional bias. Shuffling reduces positional bias, potentially utilizing more relevant evidence for the final summary, but suffers a penalty in terms of citation and evidence quality.

4.3 RQ3: Is Summary Quality Improved?

Finally, we test if learning to cite has a positive impact on summary quality. For this experiment, we measure the relevance and consistency of every summary with respect to its context and query. We again compare each base model to training on SUnsET with standard and shuffled context. Our results are presented in Table 4.

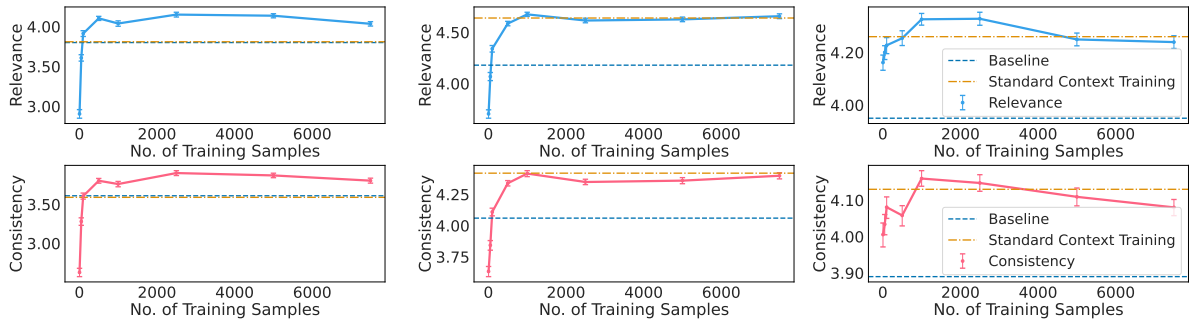
We find that **summary quality is uniformly and significantly improved by learning to cite unstructured evidence.** Larger models adapt more easily than smaller models which struggle on the single-document datasets (SQuALITY and LexAbSumm). Learning from SUnsET has an especially strong impact on multi-document datasets, while standard and shuffled training generally lead to

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



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

Figure 6: 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 7: ScholarQABench: Relevance and consistency performance vs. number of synthetic training samples.

similar gains in performance. Mistral Nemo tends to perform better with shuffled training, while for other models the results are mostly interchangeable. The selection of which approach is best then comes down to the tradeoff between evidence quality and positional bias. Ultimately though, fine-tuning on SUnsET helps reduce the performance gap in terms of summary quality with much more powerful models such as GPT 4o Mini.

Additionally, for LexAbSumm, we see a drop in performance for Llama models but gains in performance or Mixtral. Recall that LexAbSumm expresses a bias in terms of where relevant information in the summary lies. With the exception of Mixtral, shuffled training tends to mitigate this drop slightly, or lead to the best performance in the case of Nemo, which may be attributed to the reduced evidence bias of this setting.

Finally, 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 6 and Figure 7. 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. Given this, we see that adapting models to our task requires minimal data and can be performed relatively cheaply.

5 Discussion and Conclusion

Generating unstructured evidence citations alongside query-focused summaries has the potential to improve user trust and transparency in LLMs. Our study has highlighted salient challenges in this task, as well as a potential solution for them. With no intervention, these models suffer from the lost-in-the-middle problem, which we demonstrate across many settings for the case of unstructured evidence citation. They additionally struggle to generate accurate unstructured evidence from their contexts. Our proposed dataset, SUnsET, serves as a useful domain-agnostic synthetic dataset to help 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, transparent, 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. 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.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is It Really Long Context if All You Need Is Retrieval? Towards Genuinely Difficult Long Context NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 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.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

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*.

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.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models. *arXiv preprint arXiv:2408.00724*.

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 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.

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 5: Huggingface identifiers for models used in our experiments.

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.

C 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.

D Model Descriptions

Table Table 5 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}

E 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

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

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.

except title generation (temperature set to 1.2) and filtering (deterministic highest probability token output)

F 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 GPT 4o Mini as our autorater due to its high performance and low cost. This has the potential to bias some of our results, as autoraters tend to favor their own outputs. Additionally, more powerful models are available for a slightly higher cost. Therefore, we validated the robustness of GPT 4o mini as an autorater by taking a sample of 710 outputs summaries from our evaluation and re-evaluating them with DeepSeek-V3 (Liu et al., 2024a). 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 GPT 4o mini.

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.

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.

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.

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.