Context-Aware Hierarchical Merging for Long Document Summarization

Litu Ou
University of Edinburgh
lituou907@gmail.com

Mirella Lapata
University of Edinburgh
mlap@inf.ed.ac.uk

Abstract

Hierarchical Merging is a technique commonly used to summarize very long texts ($>$100K tokens) by breaking down the input into smaller sections, summarizing those sections individually, and then merging or combining those summaries into a final coherent summary. Although it helps address the limitations of large language models (LLMs) with fixed input length constraints, the recursive merging process can amplify LLM hallucinations, increasing the risk of factual inaccuracies. In this paper, we seek to mitigate hallucinations by enriching hierarchical merging with context from the source document. Specifically, we propose different approaches to contextual augmentation ranging from *replacing* intermediate summaries with relevant input context, to *refining* them while using the context as supporting evidence, and *aligning* them implicitly (via citations) to the input. Experimental results on datasets representing legal and narrative domains show that contextual augmentation consistently outperforms zero-shot and hierarchical merging baselines for the Llama 3.1 model family. Our analysis further reveals that refinement methods tend to perform best when paired with extractive summarization for identifying relevant input[1].

1 Introduction

Large language models (LLMs) have made significant strides in processing longer contexts, with models like Gemini (Team, 2024a) reportedly being able to handle up to 2 million tokens. Despite promising potential for complex real-world applications such as long-document question-answering and repository-level code understanding, most LLMs, especially open-source ones, struggle with tasks operating over long input sequences (Andryushchenko et al., 2024). A notable application we focus on in this paper is the summarization

---

[1]Our code and data are available at: https://github.com/Leonard907/CAHM
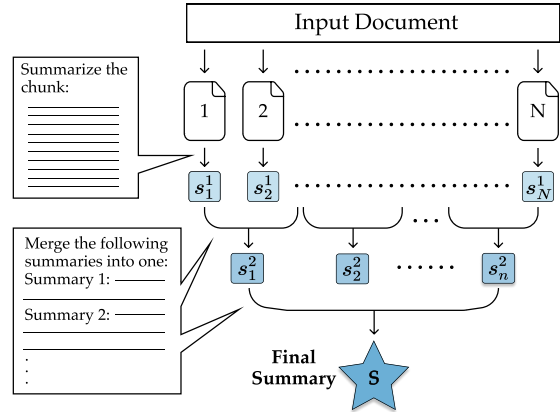

Figure 1: Illustration of hierarchical merging. The document is split into chunks and a first-level summary s_i^1 is generated for each. Second-level summaries s_j^2 are created by merging first-level summaries, and the final summary is produced by merging second-level summaries. Message boxes on the left provide examples of the prompts used for summarizing chunks and merging summaries at each level.

of very long documents in domains such as law and narrative which typically exceed input size limitations of most current models ($>$100K tokens).

Previous studies have introduced specialized techniques for summarizing very long texts, with *hierarchical merging* (Wu et al., 2021; Chang et al., 2024) being one of the most popular methods. As illustrated in Figure 1, hierarchical merging first divides an input document into chunks of fixed token length, and a summary is generated for each chunk. In subsequent stages, consecutive chunk-level summaries are concatenated and merged until they reach a pre-defined length limit, after which a new summary is generated from the merged text. Merging continues until a single final summary is produced. All summary generation steps are performed using LLMs and zero-shot prompting.

Although hierarchical merging effectively handles inputs exceeding model length limits, it cannot produce faithful summaries even when using the

most performant LLMs (Kim et al., 2024; Li et al., 2024). In this paper, we seek to mitigate hallucinations and enhance summary faithfulness by enriching hierarchical merging with *relevant context* from the source document. Access to input information during the merging process should provide models with *factual information* to generate more accurate summaries. A key question concerns how relevant context can be identified and used in the merging process. We propose three methods for selecting context from the source document: 1) *extractive summarization* identifies key sentences from the input to produce a coherent abstractive summary; 2) *retrieval* uses initial summaries as queries to retrieve relevant input passages, and 3) *attribution* generates summaries with references to input passages. Once relevant context is identified, it can either *replace* the abstractive summary or serve as *supporting evidence* for subsequent merging steps.

Experimental results with the Llama-3.1 model family, demonstrate consistent improvements (across metrics) on summary faithfulness and quality, compared to the original hierarchical merging method and zero-shot summarization. Our analysis further indicates that abstractive summaries play a crucial role in hierarchical merging by ensuring comprehensiveness, while augmenting them with relevant contexts enhances summary faithfulness. Our contributions are summarized as follows:

- We develop a novel pipeline that incorporates input context into hierarchical merging for very long document summarization.

- We propose three methods for selecting contexts: extractive summarization, retrieval, and text generation with citations.

- We analyze how the mechanism of contextual augmentation as a replacement to intermediate summaries or supporting evidence affects the faithfulness of the final summary.

2   Related Work

The literature is rife with proposals aiming to mitigate LLM difficulties with processing inputs that exceed their context lengths. A significant challenge in overcoming the context limit is addressing the quadratic computational burden of the self-attention mechanism. Previous work has attempted to reduce this cost by architectural modifications such as introducing sparse attention (Child et al., 2019; Beltagy et al., 2020), linearized attention (Katharopoulos et al., 2020), changes to positional encoding (Peng et al., 2023; Chen et al., 2023), and special-purpose fine-tuning methods which directly modify the attention mechanism (Chen et al., 2024) or incorporate discourse-level information.

Another line of work adopts a divide-and-conquer approach, splitting the long input into manageable chunks. The chunks can be processed independently (Wang et al., 2023; Bertsch et al., 2023) or through progressively merging adjacent chunks as they are processed along the transformer layers (Song et al., 2024). Wu et al. (2021) propose a dive-and-conquer approach for long document summarization in which an LLM is fine-tuned via reinforcement learning to summarize each chunk and then hierarchically merge chunk-level summaries into a final summary. Their method has since been adapted to use zero-shot prompting for summary generation (Chang et al., 2024; Kim et al., 2024), without further training (as shown in Figure 1).

Our proposal modifies hierarchical merging to consider evidence from the source document at the intermediate summary generation stage. It draws inspiration from retrieval augmented generation (RAG) techniques which have proven particularly effective in long-form question answering (Lewis et al., 2020; Izacard and Grave, 2021; Xu et al., 2024; Edge et al., 2024), as a means to integrate up-to-date information, mitigate hallucinations, and enhance response quality. Our work explores various proposals for context integration, e.g., as a replacement of intermediate summaries or additional supporting evidence, as well as retrieval augmentation methods. In addition to traditional query-based retrieval which we implement using the intermediate summaries as queries, we also turn to extractive summarization to select and rank source document sentences that best represent its content.

Various methods have been proposed for long document extractive summarization, focusing on reinforcement learning (Gu et al., 2022; Bian et al., 2023) and adapting LLMs for sentence extraction (Lu et al., 2023; Hemamou and Debiane, 2024). However, these methods were primarily evaluated on datasets with relatively short inputs, such as arXiv (Cohan et al. 2018; 5K tokens) and GovReport (Huang et al. 2021; 10K tokens), making their effectiveness on inputs exceeding 100K tokens uncertain. As we identify salient sentences within each chunk, we expect extractive methods to be able to isolate supporting evidence for ultimately generating more factual *abstractive* summaries.
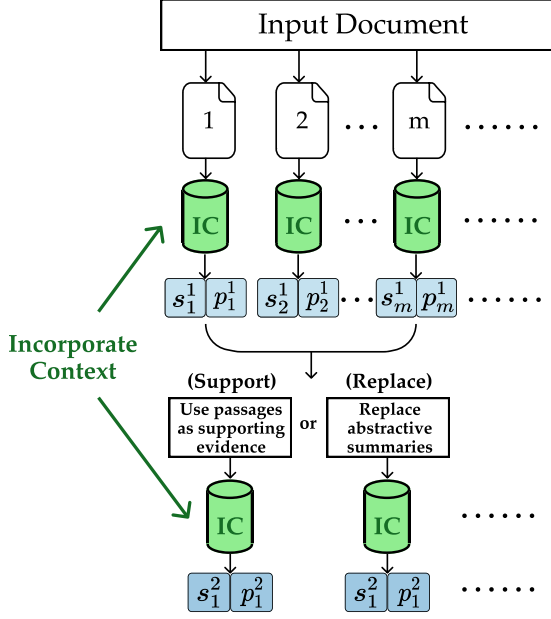
Figure 2: Illustration of proposed pipeline for the first two levels. The Incorporate Context (IC) module is applied at the first-level to obtain relevant input contexts p_i^1 alongside summaries s_i^1. At the second-level, we either use $p_1^1 \ldots p_m^1$ as supporting evidence (**Support**), or replace $s_1^1 \ldots s_m^1$ entirely with $p_1^1 \ldots p_m^1$ (**Replace**) for generating s_1^2 and subsequently obtaining p_1^2.

Our work also has connections to recent efforts in creating verifiable systems that generate responses to queries by incorporating citations to source material (Fierro et al., 2024; Nakano et al., 2022). Most existing models learn to generate citations (Menick et al., 2022; Nakano et al., 2022) or implement a post-processing step (Bohnet et al., 2022; Gao et al., 2023) where the text is first generated and then edited to be made attributable (e.g., to retrieved web content). For hierarchical merging, generating intermediate summaries with citations is advantageous as it eschews the need for additional retrieval or extraction mechanisms. However, we experimentally find that generating *accurate* citations is challenging requiring models to understand and execute complex instructions.

3 Hierarchical Merging with Context

Figure 2 illustrates our proposed hierarchical merging pipeline for the first two levels. Compared to the original method in Figure 1, at each level, an additional Incorporate Context (IC) module is used to obtain relevant contexts alongside abstractive summaries. When generating intermediate summaries for the current level, we either use the summaries from the previous level *and* relevant context

as supporting evidence (denoted as Support), or replace the summaries with their corresponding contexts (denoted as Replace). At each level, the IC module takes abstractive summaries and input contexts from the previous level to obtain relevant contexts for the current level. We make sure that input contexts have similar length to their respective abstractive summaries, so that the former can serve as a substitute for the latter.

Figure 3 provides an overview of the IC module. At the first-level, chunks from the input document are used to obtain abstractive summaries and corresponding input contexts. In Sections 3.1–3.3, we explain in detail how we propose to obtain relevant contexts. In subsequent levels, merging involves abstractive summaries (Concatenated Summaries) and their contexts (Concatenated Passages). If we choose to Replace summaries with their context, we simply substitute chunks with Concatenated Passages and repeat the first-level process. If we use input contexts as supporting evidence (Support), next level contexts are selected from Concatenated Passages, while next level summaries are generated by leveraging Concatenated Passages and Concatenated Summaries. Prompts for summary generation and merging are given in Appendix A.

To obtain contexts that are as relevant as possible, we utilize and modify three different methods which we describe as follows. The implementation of these methods in our experiments is further explained in Section 4.

3.1 Extractive Summarization: Extract

Extractive summarization selects salient sentences from the source document to form a summary. Since extractive summarizers are trained using abstractive summaries as targets (Liu and Lapata, 2019; Narayan et al., 2018; Xu and Lapata, 2022), they naturally select sentences that cover key information similar to their corresponding abstractive summaries, making them ideal input contexts for either replacement or supporting evidence. At the first-level, the IC module applies extractive summarization to input chunks. At subsequent levels, extraction operates over concatenated passages from the previous-level, ensuring selected contexts always originate from the source document.

3.2 Information Retrieval: Retrieve

Our use of information retrieval is motivated by the recent success of retrieval-augmented language models at minimizing hallucinations and improv-
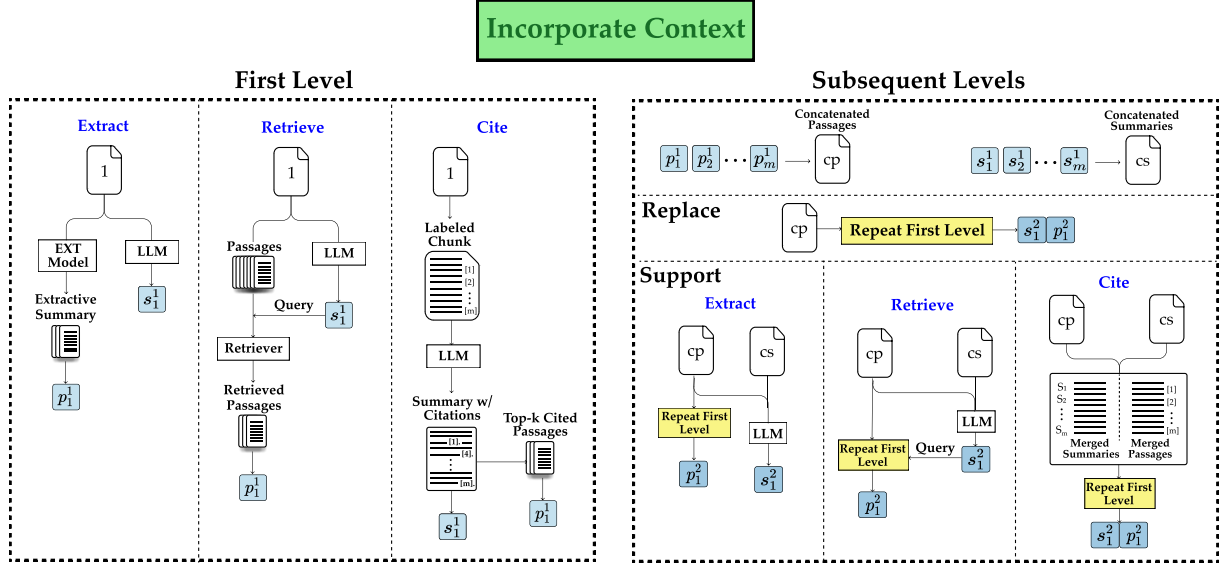
Figure 3: Incorporate Context (IC) module from Figure 2 with three context augmentation methods: Extractive Summarization (Extract), Retrieval-augmented Generation (Retrieve), and Generating text with citations (Cite).

ing the factuality of generated outputs (Shi et al., 2023; Asai et al., 2023, 2024). In the context of our summarization task, we use intermediate summaries as queries to retrieve relevant input contexts. Compared to Extract, this approach leverages the ability of recent LLMs to distill important information from the input and describe it succinctly. At the first-level, the IC module generates an abstractive summary for each chunk. We then split each chunk into passages of equal length (~100 words) and use the abstractive summary as query to retrieve top-k passages from the chunk. At subsequent levels, the retriever selects input contexts from concatenated passages at the previous-level.

3.3 Text Generation with Citations: Cite

The generation of intermediate summaries with citations as supporting evidence to input passages (Nakano et al., 2022; Fierro et al., 2024), avoids the use of information retrieval. We split the input into passages of equal length like in Section 3.2, and provide LLMs with explicit instructions to cite passages after each sentence in their summaries (see Appendix A, prompts in Tables 6 and 8). As a result, the summaries show (via their citations) which passages are used to generate which sentence. We then extract citations from the summary and rank them based on citation frequency, selecting the top-k passages as input contexts. When passages share equal citation counts, we prioritize those that enhance coverage across different input sections. Details of this selection algorithm are

given in Appendix B.

At the first-level, we add a label after each passage in square brackets with ascending numbers ($[1] \cdots [m]$; see Figure 3). A summary with citations is then generated from the labeled chunks. We employ the selection algorithm in Appendix B to identify relevant contexts, before stripping the citations to get a clean summary that will be used later. At subsequent levels, we consider two cases:

- **Replace:** We repeat the first-level process, with concatenated passages as input.

- **Support:** We instruct the LLM to generate a summary based on concatenated summaries from the previous level and concatenated passages the citations are pointing to.

4 Experimental Setup

Datasets Our experiments are conducted on two long document summarization datasets. The first dataset, Multi-LexSum (Shen et al., 2022), comprises U.S. federal court documents paired with expert-written summaries at multiple levels of granularity (long, short, and tiny summaries). For evaluation purposes, we utilize the long summaries as a reference and exclude documents with fewer than 100K tokens. The second dataset, SuperSummary, is a novel book summarization corpus that we constructed using literary works published after 2022. The corresponding expert-written plot summaries were obtained from SuperSummary[2]

---

[2]https://www.supersummary.com/

through a paid subscription service. Similarly, we employ these plot summaries as gold summaries for evaluation and exclude documents containing fewer than 100K tokens. Table 1 presents the statistics of both datasets.

Implementation Details Our experimental evaluation was conducted using the Llama 3.1 model family (Team, 2024b), which features a context length of 128K tokens and comprises two variants with 8B and 70B parameters, respectively. We use the GPTQ-INT4 quantized 70B model (Frantar et al., 2023) in our experiments for inference efficiency. For all hierarchical merging systems, the chunk size and maximum merging context length were configured to 8K tokens, ensuring sufficient merging operations for meaningful analysis of different hierarchical merging variants.

As mentioned earlier, we propose three methods for incorporating context, i.e., Extract, Retrieve, and Cite. We use MemSum (Gu et al., 2022), a state-of-the-art extractive summarizer trained with reinforcement learning, as our backbone model for Extract. We fine-tune MemSum on the official Multi-LexSum training split. For SuperSummary, we use the training split of the publicly accessible BOOKSUM dataset (Kryscinski et al., 2022). To prepare these datasets for MemSum fine-tuning, each document was segmented into 8K-token chunks, from which 1,000 chunks (per dataset) were randomly selected as training inputs. Subsequently, we use the full Llama-3.1 70B model to generate summaries for each chunk via zero-shot prompting, which serve as reference summaries. The fine-tuning process was executed for 10 epochs with a learning rate of 1e-4. To maintain consistency with the average summary length in the evaluation datasets, the maximum number of extracted sentences was set to 20.

For Retrieve, we split input documents into uniform length passages with 100 tokens each. To optimize compute efficiency, we utilize the BM25 retriever to select passages using intermediate summaries as queries. The top-k parameter for selecting passages is set so as to ensure concatenated top-k passages have lengths similar to the average summary length in the evaluation datasets. Similar to Retrieve, we set the passage length to 100 tokens each for Cite. The top-k passage selection process is described in Section B.

Model Comparisons As described in Section 3, the proposed hierarchical merging pipeline of-

	Multi-LexSum	SuperSummary
# examples	151	194
input tokens	156,447	133,246
summary tokens	1,147	1,180
domain	legal	narrative

Table 1: Statistics of the datasets used in this paper. We report the number (#) of examples per dataset, and the average number of input and summary tokens.

fers two approaches for incorporating context beyond the first level: either via replacing abstractive summaries with corresponding input contexts (Replace) or via using input contexts as supporting evidence (Support). This together with the three context selection methods (Extract, Retrieve, Cite) yields six distinct model invitations: Extract-Replace, Extract-Support, Retrieve-Replace, Retrieve-Support, Cite-Replace, and Cite-Support.

We evaluate these models against three baseline approaches: (a) Zero-shot summarization (Zeroshot), which generates summaries through zero-shot prompting after truncating the input to the maximum context length; (b) the original hierarchical merging method (HMerge), as illustrated in Figure 1; and an adaptation thereof which incorporates citation generation (Cite-HMerge). Each intermediate summary is stripped off its citations and re-labeled together with other merged summaries at the merging step.

Evaluation Metrics We use three automatic metrics to evaluate the *faithfulness* of our summaries:

SUMMAC (Laban et al., 2021) is a specialized metric developed for assessing the semantic consistency between generated summaries and input contexts. Its architecture is particularly effective for long-document summarization, as it implements natural language inference at the sentence level and employs a convolutional layer to aggregate results into a single final consistency score.

AlignScore (Zha et al., 2023) evaluates the information alignment between generated texts and the corresponding source texts through a unified text-to-text alignment function. While SUMMAC is specifically optimized for summarization tasks, AlignScore is a more general metric and demonstrates robust performance across various applications, including summarization and fact verification. This metric implements chunk-based processing to handle long documents, similar to SUM-

MAC's approach.

PRISMA (Mahon and Lapata, 2024), is a recent summarization metric built upon *FActScore* (Min et al., 2023). This metric decomposes generated texts into atomic facts—discrete statements containing singular pieces of information—and evaluates factual consistency between generated and reference summaries at this granular level. The methodology involves a two-step process: first, extracting atomic facts from both generated and reference summaries, then computing fact precision by computing the number of generated summary facts entailed by the reference summary, and fact recall by computing the number of reference summary facts entailed by the generated summary. The final metric is derived from the harmonic mean of these precision and recall measurements. Unlike SUMMAC and AlignScore, PRISMA evaluates summary quality through comparison with reference summaries, allowing it to better capture whether the generated summary preserves the same key information that human annotators deemed important in their reference summaries. We use the GPT-4o-mini model for both fact extraction and entailment computation, incorporating in-context learning examples to ensure the quality of atomic fact extraction, following Mahon and Lapata (2024).

In addition to faithfulness metrics, we report two standard metrics that assess the *quality* of generated summaries against reference summaries: ROUGE (Lin, 2004) as the geometric mean of ROUGE-1, ROUGE-2, and ROUGE-L (Shaham et al., 2023) and BERTScore (He et al., 2021) (we use the `deberta-xlarge-mnli` model and report F1). Overall, we employ three input-based metrics (SUMMAC and AlignScore), which evaluate against the source text and three reference-based metrics (PRISMA, ROUGE, and BERTScore), which compare against reference summaries.

5 Results

Tables 2 and 3 present our results for the Llama-3.1 8B and 70B models, respectively. Hierarchical merging with incorporated context overwhelmingly outperforms comparison systems without supporting evidence (see first block in the tables). For each metric (and dataset) the best performing model is a variant of hierarchical merging with context (the only exception is BERTScore on SuperSummary for Cite-HMerge using the 8B model).

Extract-Support is the best model. Among the proposed approaches, Extract-Support demonstrates superior performance, achieving the highest average score across all metrics. Cite emerges as the least effective context selection method, showing superior performance only in AlignScore with minor differences to second best results. When examining the impact of model size, scaling to 70B parameters yields moderate improvements on reference-based metrics, although gains on input-based metrics are less substantial. Additionally, our analysis indicates that models consistently produce lower scores on SuperSummary compared to Multi-LexSum, suggesting that book summarization poses greater challenges than legal document summarization.

Replace models show big improvements according to input-based metrics. Examining input-based metrics more closely, both SummaC and AlignScore show significant improvements when input passages are incorporated via Replace. Specifically, for the 8B model, while Support achieves similar results to baselines, Replace consistently demonstrates superior performance, with Retrieve-Replace and Cite-Replace achieving approximately 10-points improvements in AlignScore on SuperSummary compared to other methods. The same patterns are again observed in the 70B model results, where Retrieve-Replace excels in SummaC scores and Cite-Replace leads in AlignScore metrics. These results align with our expectations, given that both metrics evaluate performance against input texts, and Replace explicitly leverages input context in summary generation.

Support models are better summarizers according to reference-based metrics. Reference-based metrics reveal patterns that stand in sharp contrast to those observed with input-based metrics. In both 8B and 70B models, Extract-Support and Retrieve-Support consistently outperform baselines. However, Cite-Support shows comparable or slightly worse performance, suggesting that while input contexts can enhance summary faithfulness, their effectiveness depends heavily on the method's ability to identify and utilize relevant segments. Compared to Support methods, Replace methods show significantly inferior performance across both model scales, with particularly pronounced differences on SuperSummary, where Replace methods lag behind Support by approximately 20 points across all three context selection methods in 70B

6

Multi-LexSum

Models	ROUGE	BScore	SUMMAC	AScore	PRISMA
Zero-shot	24.6	60.6	40.9	79.3	42.7
HMerge	24.4	61.7	41.1	72.9	45.5
Cite-HMerge	23.2	62.4	43.3	82.0	43.0
Retrieve-R	23.4	61.5	**47.0**	**84.6**	41.8
Retrieve-S	<u>24.9</u>	<u>62.6</u>	41.2	79.4	<u>46.1</u>
Extract-R	24.1	61.6	43.2	79.4	41.9
Extract-S	**25.8**	**63.3**	40.6	79.4	**47.5**
Cite-R	20.4	59.2	<u>46.8</u>	<u>82.3</u>	36.3
Cite-S	23.0	61.5	42.7	81.0	43.0

Multi-LexSum

Models	ROUGE	BScore	SUMMAC	AScore	PRISMA
Zero-shot	23.6	60.7	43.5	77.6	41.5
HMerge	<u>26.7</u>	<u>64.3</u>	43.4	76.3	48.2
Cite-HMerge	25.6	64.0	43.1	82.7	47.3
Retrieve-R	24.7	62.1	<u>47.9</u>	80.3	43.8
Retrieve-S	26.6	**66.1**	44.5	78.8	**49.8**
Extract-R	24.8	60.4	43.4	79.9	42.7
Extract-S	**27.6**	64.1	43.2	79.0	<u>49.7</u>
Cite-R	22.5	61.9	**51.6**	**85.8**	40.6
Cite-S	25.4	62.5	44.4	<u>83.9</u>	46.8

SuperSummary

Models	ROUGE	BScore	SUMMAC	AScore	PRISMA
Zero-shot	18.1	57.8	38.8	60.7	33.2
HM	<u>20.3</u>	63.3	38.7	55.8	37.8
Cite-HMerge	19.8	**66.0**	38.1	59.1	35.1
Retrieve-R	16.5	60.0	**47.6**	<u>70.2</u>	23.3
Retrieve-S	19.9	<u>65.4</u>	38.6	56.4	<u>38.4</u>
Extract-R	16.8	58.1	<u>41.6</u>	61.1	22.3
Extract-S	**21.2**	63.7	38.3	57.2	**39.2**
Cite-R	15.6	58.1	40.4	**71.8**	20.3
Cite-S	19.3	62.3	37.9	55.1	34.6

SuperSummary

Models	ROUGE	BScore	SUMMAC	AScore	PRISMA
Zero-shot	18.9	58.1	38.4	57.9	35.2
HMerge	21.8	65.4	39.1	57.8	42.2
Cite-HMerge	20.9	<u>66.7</u>	38.4	59.2	38.8
Retrieve-R	16.9	58.9	**44.4**	<u>66.7</u>	25.4
Retrieve-S	<u>22.4</u>	65.5	38.7	59.3	<u>44.3</u>
Extract-R	17.2	59.0	41.6	59.7	23.5
Extract-S	**22.9**	**67.2**	37.6	58.5	**45.6**
Cite-R	15.5	59.9	<u>43.6</u>	**78.4**	21.8
Cite-S	20.6	65.3	37.8	57.0	40.8

Table 2: Results with Llama-3.1-8B-Instruct model. We report the geometric mean of ROUGE-1, ROUGE-2, and ROUGEL-L (ROUGE). BScore and AScore are shorthands for BERTScore and AlignScore, respectively. Labels -R and -S, refer to Replace and Support context incorporation methods. Best results for each metric are highlighted in **bold**, whereas the second best results for each metric are <u>underlined</u>.

Table 3: Results for the Llama-3.1-70B-Instruct model. We report the geometric mean of ROUGE-1, ROUGE-2, and ROUGEL-L (ROUGE). BScore and AScore are shorthands for BERTScore and AlignScore, respectively. Labels -R and -S, refer to Replace and Support context incorporation methods Best results for each metric are highlighted in **bold**, while the second best results for each metric are <u>underlined</u>.

models. This performance gap is further corroborated by both ROUGE and BERTScore metrics. In summary, results with reference-based metrics strongly suggest that methods incorporating abstractive summaries during hierarchical merging (as in Support and baseline approaches) are better at maintaining comprehensive coverage compared to methods that rely solely on input contexts for final summary generation (Replace approaches), which may become overly specific and miss broader aspects of the source material.

Adding more contexts yields better summaries for Replace. Our results in Tables 2 and 3 suggest that incorporating contexts through Replace degrades performance across reference-based metrics. Manual inspection of Replace summaries further showed that they tend to emphasize different key events compared to Support or baseline methods, often prioritizing peripheral details such as ca-

sual dialogue or scene descriptions. Despite having comparable lengths, final-stage inputs from Support and baseline methods have higher information density compared to Replace. Specifically, Replace contexts typically elaborate on a single key event extensively, while abstractive summaries convey the same event more concisely, allowing space for additional key events.

Building on these observations, we investigate whether Replace summaries can be improved by increasing the maximum merging context length, thereby incorporating more information while merging the same number of summaries. Specifically, we extend the maximum merging context length from 8K to 16K and 32K tokens. We maintain the same retrieval units but select more passages that fit within this increased context length window, thereby generating next-level summaries with greater information density. We experiment

Context Length	Multi-LexSum		SuperSummary	
8K	41.8	—	23.3	—
16K	43.7	+1.9	26.5	+3.2
32K	44.1	+2.3	27.9	+4.6

Table 4: PRISMA results for Retrieve-Replace with different maximum merging context lengths Llama-3.1-8B-Instruct. Gains compared to 8K context length are shown in red.

Models	Correct	Incorrect	Not Present
Extract-Support	72.7%	18.2%	9.1%
Cite-HMerge	60.0%	20.0%	20.0%
Zero-shot	60.0%	20.0%	20.0%
HMerge	59.1%	27.3%	13.6%
Extract-Replace	48.8%	23.3%	27.9%

Table 5: Proportion of Correct, Incorrect, and Not Present claims for summaries generated by Extract-Support, Extract-Replace, and comparison baselines using Llama-3.1-70B-Instruct (GPTQ-INT4). Models summarize the book *Things we Hide from the Light* (by Lucy Score) from the SuperSummary dataset.

with context extension on the 8B model using both datasets, employing Retrieve-Replace as the hierarchical merging method. The results, shown in Table 4, indicate that using 16K and 32K tokens achieves higher PRISMA scores on both datasets, with more substantial gains on SuperSummary compared to Multi-LexSum. However, even with the increased threshold of 32K tokens, performance still falls short of the results achieved by Retrieve-Support, demonstrating that the use of abstractive summaries is critical for achieving better coverage.

Manual analysis confirms Extract-Support generates the most faithful summaries. Our results suggest that input- and reference-based metrics tend to favor different methodological approaches. To evaluate metric reliability, we analyzed book summaries generated by the 70B model. Specifically, we decomposed output summaries for the book *Things we Hide from the Light* (by Lucy Score) into atomic claims and verified whether they are supported by the source. We followed a simplified version of the annotation guidelines proposed in Kim et al. (2024). Claims were classified as 'Correct' if fully supported by the book, 'Incorrect' if parts of the claim contradict the source, or 'Not Present' if parts of the claim are absent from the source material.

Table 5 presents the results of our analysis for Extract-Support, Extract-Replace, and the three comparison baselines (Cite-HMerge, Zero-Shot, and HMerge). The summaries and claim annotations can be found in Appendix D and E, respectively. We observe that Extract-Support summaries contain the highest proportion of Correct claims (10% higher compared to baseline methods). When comparing claims from Extract-Support and the original HMerge, we find that while the summaries maintain overall similar structure, there are notable differences in the selection of key events. These differences can be attributed to the incorporation of supporting evidence, which may influence the

prioritization of events mentioned in such contexts.

In contrast, Extract-Replace performs below baseline levels, by more than 10%. This poor performance may be explained by inherent challenges in extractive summarization, where the merging of input contexts can introduce coherence issues that potentially confuse LLMs, leading to unfaithful summary generation (Zhang et al., 2023). The relative performance of these methods (Extract-Support being most factual and Extract-Replace being least factual) aligns with automatic results using reference-based metrics such as PRISMA (see last column in Tables 2 and 3). Taken together, human-based and automatic evaluation results suggest that comprehensive source coverage contributes to enhanced faithfulness in the final summary.

6 Conclusion

In this paper, we examined how input contexts can be integrated into hierarchical merging so as to produce more faithful summaries for very long source documents (>100K tokens).

We proposed different approaches to contextual augmentation ranging from *replacing* intermediate summaries with relevant input context, to *refining* them while using the context as supporting evidence, and implicitly *aligning* summary sentences (via citations) to input passages. Our findings indicate that contextual augmentation leads to more factual summaries. However, input contexts alone can distract from important information, emphasizing peripheral details, and thus work best to support rather than substitute abstractive summaries that capture a broader perspective of the source content.

In the future, we would like to learn when to perform context augmentation based on the quality of the intermediate summary (i.e., calling context augmentation as a tool) and to explore trade-offs

between context length and summary quality.

7 Limitations

Large-scale Human Evaluation We verify automatic evaluation results with human annotations performed on model summaries corresponding to one book. Human evaluation results could be enhanced by performing experiments on multiple books, using a bigger annotator pool. There are well-documented challenges associated with human-based evaluations of summaries based on very long input documents (Chang et al., 2024; Krishna et al., 2023; Xu et al., 2023; Kim et al., 2024). In our case, hiring multiple annotators to read and understand books (or legal documents) would have been prohibitively expensive and time-consuming. Kim et al. (2024) report annotation costs in the range of $200 to $250 per book, requiring 10 hours of work (excluding the time it takes to read the book).

Efficient Run-time for Support Support summaries are generated using input contexts as supporting evidence. This results in longer inputs for LLMs during the merging phase compared to approaches that do not include supporting evidence. Future research could explore more efficient ways of incorporating contextual information into the summarization process, by selectively including context, e.g., based on positional information or knowledge about the structure of the input.

Limited Dataset Domain This work focuses on the specific task of full-text summarization for very long documents (>100K tokens). This constraint significantly limits the available datasets, restricting both the domains and samples we can use to test our proposed pipeline. Although we identified other potentially suitable datasets, such as LFOSum (Nayeem and Rafiei, 2024), which focuses on opinion summarization with an average input length of 207K tokens, these were not open-sourced. Overall, there is a pressing need to curate more high-quality summarization datasets containing inputs exceeding 100K tokens across diverse domains.

The Role of Context Our results so far have shown that incorporating context generally improves the faithfulness of the generated summaries. More fine-grained analysis is required to establish how contexts improve the quality of the intermediate summaries as well as the end result. For example, it would be interesting to study how to dynamically select contexts to increase efficiency and avoid unnecessary distraction from using contexts at merges that don't need them.

Acknowledgments

References

Georgy Andryushchenko, Vladimir Ivanov, Vladimir Makharev, Elizaveta Tukhtina, and Aidar Valeev. 2024. Leveraging large language models in code question answering: Baselines and issues. *Preprint*, arXiv:2411.03012.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *Preprint*, arXiv:2403.03187.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *arXiv preprint arXiv:2305.01625*.

Junyi Bian, Xiaodi Huang, Hong Zhou, and Shanfeng Zhu. 2023. Gosum: Extractive summarization of long documents by reinforcement learning and graph organized discourse state. *Preprint*, arXiv:2211.10247.

Bernd Bohnet, Vinh Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, Lierni Sestorain Saralegui, William Weston Cohen, Michael Collins, Dipanjan Das, Don Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *Preprint*, arXiv:2306.15595.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. Longlora: Efficient fine-tuning of long-context large language models. *Preprint*, arXiv:2309.12307.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *Preprint*, arXiv:1904.10509.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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. *Preprint*, arXiv:2404.16130.

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. *Preprint*, arXiv:2404.03381.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Nianlong Gu, Elliott Ash, and Richard Hahnloser. 2022. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

L'eo Hemamou and Mehdi Debiane. 2024. Scaling up summarization: Leveraging large language models for long text extractive summarization. *ArXiv*, abs/2408.15801.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. *Preprint*, arXiv:2404.01261.

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

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Preprint*, arXiv:2111.09525.

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

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Guang Lu, Sylvia B. Larcher, and Tu-Anh Tran. 2023. Hybrid long document summarization using c2f-far and chatgpt: A practical study. *ArXiv*, abs/2306.01169.

Louis Mahon and Mirella Lapata. 2024. A modular approach for multimodal summarization of TV shows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8272–8291, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2024. Lfosum: Summarizing long-form opinions with large language models. *Preprint*, arXiv:2410.13037.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *Preprint*, arXiv:2309.00071.

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 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *CoRR*, abs/2206.10883.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *Preprint*, arXiv:2301.12652.

Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, and Jinwoo Shin. 2024. Hierarchical context merging: Better long context understanding for pre-trained llms. *Preprint*, arXiv:2404.10308.

Google Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Meta Llama 3 Team. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. *Preprint*, arXiv:2306.07174.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Francis Christiano. 2021. Recursively summarizing books with human feedback. *ArXiv*, abs/2109.10862.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. *Preprint*, arXiv:2310.03025.

Yumo Xu and Mirella Lapata. 2022. Text summarization with oracle expectation. *Preprint*, arXiv:2209.12714.

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

Shiyue Zhang, David Wan, and Mohit Bansal. 2023. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2153–2174, Toronto, Canada. Association for Computational Linguistics.

A   Prompts

Table 6 contains first-level prompt templates, for abstractive summaries and summaries with citations. Table 7 shows subsequent-level templates for the original hierarchical merging (HMerge) and our adaptation with citations. Table 8 presents subsequent-level templates for Extract-Support, Retrieve-Support, and Cite-Support.

B   Algorithm for Ranking Citations

Algorithm 1 describes how we select passages based on citation labels extracted from the summary. First, we build a dictionary with each label being the key, and the number of times it appears in the summary its value. We then append all passages that belong to the most frequent label, remove this label from the dictionary, and repeat this process until all passages are appended or number of passages to be selected (denoted as k) is exceeded. If no passages are left or we have appended exactly k passages, we return the selected passages. Otherwise, we select passages for the next label to fill the remaining vacancies.

When filling the remaining vacancies, we aim to maximize the coverage of the input. We first divide the input text into $k-1$ sections of equal length, so that ideally the selected passages should cover all sections. For each passage that has been already selected, we find its closest section based on its position in the input text and remove that section from consideration. Then, for each remaining candidate passage (those belonging to the next most frequent label), we calculate its distance to each remaining section. We iteratively select passages that have the minimum distance to any remaining uncovered section, removing that section from consideration after each selection. This process continues until we have selected exactly k passages in total. Overall, This approach ensures that the selected passages not only represent the most frequently cited content (through label frequency) but also provide broad coverage across different portions of the source text, avoiding redundant selection from the same regions.

C   Additional Results on Fact-Precision and Fact-Recall for PRISMA

Tables 9 and 10 present results for Fact-Precision and Fact-Recall for the 8B and 70B Llama models, respectively. In general, we see that Extract-Support has the best overall performance, achieving the highest scores most frequently. Retrieve-Support trails closely behind with multiple highest and second highest appearances. We also observe that the original HMerge method shows strong performance at FACT-PREC, but falls behind on FACT-REC which leads to worse PRISMA scores.

D   Examples of Model Output

This section contains output summaries for different models for the book *Things We Hide from the Light* (by Lucy Score). Table 11 contains the Reference summary, whereas Tables 12, 13, 14, 15, and 16 respectively present summaries corresponding to the Zero-shot model, the original HMerge method, Cite-HMerge our adaptation of HMerge with citations, and the proposed Extract-Support summary and Extract-Retrieve methods.

E   Manual Analysis of Summary Facts

Tables 17, 18, 19, 20 and 21 respectively show our manual annotations for the Zero-shot summary, the original HMerge summary and its adaptation with citations Cite-HMerge, the Extract-Support and Extract-Replace summaries. "✓" means the statement is supported by the source book text. "✗" means the statement or parts thereof are contradicted by the source book text. "–" means that the statement or parts thereof are not present in the source book text. We implemented the following annotation procedure:

- If every part of the statement is entailed by the book source, we classify the statement as Correct ("✓").

- Otherwise, we check whether some parts of the statement contradict the book source and classify the statement as Incorrect ("✗"). A piece of text *contradicts* the source if it has information is incompatible or contradictory with explicit statements or facts presented in the book source.

- If none of the above apply, we classify the statement as Not Present ("—"), since it expresses information that is neither directly sup-

Algorithm 1 Passage Selection with Coverage Constraints

attrTexts ← List of input passages
response ← The summary w/ citations
k ← No. of passages to select

allLabels ← EXTRACTLABELS(response)
labelCounts ← COUNTLABELFREQUENCIES(allLabels)   ▷ dict in format {label: count}
SORT(labelCounts, key=count, order=descending)

selectedPassages ← []
labelToPick ← None
for all (label, count) ∈ labelCounts **do**
    if count + LEN(selectedPassages) > k **then**
        labelToSelect ← label
        break
    else
        selectedPassages += GETPASSAGESBYLABEL(attrTexts, label)
    end if
end for

if labelToSelect == None **then**
    return selectedPassages
else
    remainingPassages ← GETPASSAGESBYLABEL(attrTexts, labelToPick)
    remainingPassageIndices ← INDEXOF(remainingPassages, attrTexts)
end if

coverageSections ← $[\frac{2i+1}{2k}$ for i ∈ $0 \cdots k-1]$
for all passage ∈ selectedPassages **do**
    pos ← INDEXOF(passage, attrTexts)
    closestSection ← FINDCLOSESTSECTION(coverageSections, pos)
    coverageSections ← coverageSections \ {closestSection}
end for

for all passage, passageIndex ∈ remainingPassage, remainingPassageIndices **do**
    closestSection ← FINDCLOSESTSECTION(coverageSections, passageIndex)
    distToClosestSection ← DISTTOCLOSESTSECTION(coverageSections, passageIndex)
    if distToClosestSection is minimum **then**
        selectedPassages ← selectedPassages ∪ {passage}
        coverageSections ← coverageSections \ {closestSection}
    end if
    if LEN(selectedPassages) == k **then**
        return selectedPassages
    end if
end for

ported or contradicted by the source (and as a result we cannot make a definitive judgment about its validity).

First-level Chunk Summary

```
Below is a document:
---
{}
---
Write a summary containing all
key information.  There should be
no explicit mention of words like
"document"  or  "summary"  in  the
summary.
```

First-level Chunk Summary w/Citations

```
Below  is  a  document  with  each
paragraph  assigned  to  a  label  at
the end ([n]) and separated by line
breaks:
---
{}
---
Write  a  summary  containing  all
key information.  There should be
no explicit mention of words like
"document"  or  "summary"  in  the
summary. After each summary sentence,
you should assign a label to that
sentence  showing  which  paragraph
in  the  document  it  corresponds  to.
Specifically,  follow  the  format
below:

<sentence 1>. [n] <sentence 2>. [m]
...
```

Table 6: First-level prompt templates for generating intermediate summaries without and with citations.

Subsequent-level Merging: HMerge

```
Below are several summaries of
different parts of a document:
---
{}
---
Merge the given summaries into
one single summary containing all
key information. There should be
no explicit mention of words like
"document" and "summary" in the
summary.
```

Subsequent-level Merging: CiteHMerge

```
Below are several summaries of
different parts of a document with
each summary having sentences with
labels at the end ([1], [2], ...)
and separated by line breaks:
---
{}
---
Merge the given summaries into
one single summary containing all
key information. There should be
no explicit mention of words like
"document" and "summary" in the
summary. After each summary sentence,
you should assign a label to that
sentence showing which paragraph
in the document it corresponds to.
Specifically, follow the format
below:

<sentence 1>. [n] <sentence 2>.
[m] ...
```

Table 7: Subsequent-level prompt templates for Original-HM and Cite-HM.

Subsequent-level Merging: Extract/Retrieve Support

```
Below are several summaries of different parts of a document:
---
{}
---
Below are the supporting contexts of the previously shown summaries:
---
{}
---
Merge the given summaries into one single summary containing all key information
and use the supporting contexts to make sure the merged summary contains no
factual errors. The gist of the summary should be based solely on the given
summaries, while the supporting contexts should be used for proofreading only.
There should be no explicit mention of words like "document", "context" or
"summary" in the summary.
```

Subsequent-level Merging: Cite-Support

```
Below are several summaries of different parts of a document:
---
{}
---
Below are the supporting contexts of the previously shown summaries, with each
context assigned to a label at the end ([n]) and separated by line breaks:
---
{}
---
Merge the given summaries into one single summary containing all key information
and use the supporting contexts to make sure the merged summary contains no
factual errors. The gist of the summary should be based solely on the given
summaries, while the supporting contexts should be used for proofreading only.
There should be no explicit mention of words like "document", "context" or
"summary" in the summary. After each summary sentence, you should assign a
label to that sentence showing which supporting context it corresponds to.
Specifically, follow the format below:

<sentence 1>. [n] <sentence 2>. [m] ...
```

Table 8: Subsequent-level prompt templates for Support methods.

Multi-LexSum	Fact-Prec	Fact-Rec
Zero-shot	50.3	37.1
HMerge	55.1	38.8
Cite-HMerge	51.7	36.8
Retrieve-Replace	<u>56.1</u>	33.3
Retrieve-Support	55.2	<u>39.6</u>
Extract-Replace	48.2	37.1
Extract-Support	**56.3**	**41.1**
Cite-Replace	40.8	32.7
Cite-Support	50.4	37.5

SuperSummary	Fact-Prec	Fact-Rec
Zero-shot	<u>60.1</u>	22.9
HMerge	**61.5**	27.3
Cite-HMerge	45.6	28.9
Retrieve-Replace	29.0	19.5
Retrieve-Support	54.5	<u>29.6</u>
Extract-Replace	33.8	16.7
Extract-Support	56.0	**30.1**
Cite-Replace	29.2	15.6
Cite-Support	46.4	27.7

Table 9: Fact-Prec and Fact-Rec Results for the Llama-3.1-8B-Instruct model. Fact-Prec and Fact-Rec refer to fact precision and recall obtained while computing PRISMA. Best results for each metric are highlighted in **bold**, while the second best results for each metric are <u>underlined</u>.

Multi-LexSum	Fact-Prec	Fact-Rec
Zero-shot	46.7	37.3
HMerge	<u>57.5</u>	41.5
Cite-HMerge	54.9	41.6
Retrieve-Replace	57.1	35.6
Retrieve-Support	57.1	**44.1**
Extract-Replace	50.6	36.9
Extract-Support	**57.8**	<u>43.7</u>
Cite-Replace	46.7	35.9
Cite-Support	54.6	41.0

SuperSummary	Fact-Prec	Fact-Rec
Zero-shot	55.8	25.8
HMerge	<u>59.4</u>	32.7
Cite-HMerge	49.9	31.7
Retrieve-Replace	40.7	18.5
Retrieve-Support	**60.6**	<u>34.9</u>
Extract-Replace	45.1	15.9
Extract-Support	57.2	**37.9**
Cite-Replace	21.5	22.1
Cite-Support	52.7	33.3

Table 10: Fact-Prec and Fact-Rec Results for the Llama-3.1-70B-Instruct (GPTQ-INT4) model. Fact-Prec and Fact-Rec refer to fact precision and recall obtained while computing PRISMA. Best results for each metric are highlighted in **bold**, while the second best results for each metric are <u>underlined</u>.

Reference Summary

It has been months since Nash Morgan, Police Chief of Knockemout, Virginia, was shot attempting to break up a stolen car ring run by Duncan Hugo, the son of an underworld kingpin in nearby Washington. Even now, FBI special agents shadow Nash to protect him against a possible reprisal by the crime family. Nash has post-traumatic stress disorder (PTSD) and cannot recall the shooting. Despite counseling and anxiety medication, Nash feels isolated and paranoid, and is prone to anxiety attacks.

Angelina Solavita, an insurance investigator known for tracking down stolen property, moves into the apartment next door to Nash. She is the ex-lover of Nashs older brother, Knox. Wrestling with her own kind of PTSD (she nearly died from a heart defect when she was 15), Lina hesitates to share with Nash that she is in town to recover a vintage Porsche stolen by the same crime ring responsible for his shooting. Although they are immediately attracted to each other, Lina sees the hunky Nash more as a fling given her nomadic lifestyle. Meanwhile, Nash, who has trauma because of the shooting, yearns for the stability of a relationship.

At work, Nash deals with Officer Tate Dilton, a rookie cop who has earned a string of harassment complaints. The department is in legal trouble after Dilton pulled over a Black couple driving a luxury car on the false suspicion (prompted by racial profiling) that the car must be stolen. The couple is now contemplating suing the department. Nash puts Dilton on paid leave.

Lina gathers evidence about the missing Porsche. As she digs, she begins to suspect the thug who shot Nash may be part of the same crime organization that stole the sports car. As Nashs and Linas investigations progress, the two act on their attraction. They share a passionate kiss, flirt, and, when Nash experiences a panic attack, they spend a chaste night together. Gradually, they open up to each other about their near-death experiences and their feelings of vulnerability. Lina tells Nash why she is in Knockemout. When the two finally have sex, it is life-altering in its intimacy and its intensity. Lina takes a reluctant Nash skydiving, and as the two fall to earth harnessed to each other, they both feel the profound pull of love.

When a drunken Tate Dilton and his beer buddies harass Lina and her friends during a town Halloween party, Nash intervenes alongside his brother and several friends. A fight breaks out. The following morning, Nash relieves Tate of his duties.

Using a network of friends, Lina and Nash coordinate their efforts to identify who shot Nash. Solving that mystery will lead them to Duncan Hugo and the stolen Porsche. On a hunch from Waylay, Nashs 12-year-old niece who was abducted by the car thieves the night Nash was shot, Lina turns her attention to identifying a shadowy man who has been watching her. Waylay recalls that on the night of Nashs shooting, one of the thugs ate wrapped candy; Lina notices that the man she keeps seeing purchases bags of the same candy at a convenience store.

Nashs network of friends digging into the Hugo criminal organization learns that Duncan wants to cooperate with federal agents in return for eliminating his own father so that he can take over the family operation. Duncan has never left Knockemout, which means that Nashs shooting was a botched effort by an unnamed underling, but the question of who and where Duncan is hiding furthers the mystery.

One day, after rushing to the scene of nasty car accident, Nash encounters his estranged father, who is recovering from misusing alcohol after Nashs mother died. Nash feels overwhelmed, helpless, and unable to protect anybody. Without telling her the reason, Nash breaks it off with Lina.

One day, two thugsone of them the candy mansnatch Lina and drive her to a barn at an abandoned horse farm outside of Knockemout. Lina is stunned to be met by Tate Dilton, who turns out to be one of Duncan Hugos henchmen.

Duncan himself appears. Ever resourceful, Lina slips her restraints and surprises Duncan and Tate as they play a video game. She snatches the console, eludes them both, and hides in one of the stalls in the barn. On a hunch, she uses the consoles IP signal to contact Waylay, who in turn contacts Nash. Within minutes, Nash and his friends swarm the farm. Nash suddenly remembers that it was Tate Dilton who shot him that night. In a tense showdown, Tate pulls his gun, and then Nash shoots him dead. Duncan Hugo is apprehended. Lina and Nash fall into each others arms.

Days later, at his brothers wedding, Nash and Lina commit to each other. Nash gives Lina an engagement ring, and the two talk happily about the future they will have together. Five years later, Lina and Nash, now married, await the birth of twins.

Table 11: Reference summary.

Zero-shot Summary

Lina Solavita, an insurance investigator, moves to Knockemout, a small town in northern Virginia, to find a 1948 Porsche 356 convertible that was stolen from a client. She is staying in an apartment next to Nash Morgan, the chief of police, who is still recovering from a shooting that left him with bullet wounds and memory loss.

Nash is struggling to cope with his new reality and is haunted by the sound of crunching and a sizzle or crunch that echoes the moment he was shot. Lina is initially attracted to Nash, but their relationship is complicated by his vulnerability and her job, which requires her to keep secrets. She is also dealing with her own emotional baggage, including a heart condition that almost killed her at 15.

Meanwhile, a U.S. marshal, Nolan Graham, is watching Nash due to a threat, and Lina is caught in the middle. As Lina and Nash spend more time together, they develop a strong connection, but their differences and secrets threaten to tear them apart. Lina's past is slowly revealed, including her history with Nolan and her job as an insurance investigator. Nash is torn between his growing feelings for Lina and his need for honesty and trust.

The story takes a dark turn when Lina's investigation puts her in danger, and Nash is forced to confront his own vulnerabilities. The two are eventually kidnapped by a group of bikers, who are revealed to be working for Grim, a motorcycle club leader. Grim reveals that Duncan Hugo, the man who shot Nash, is still in town and planning to take over his father's criminal empire.

As the story unfolds, Lina and Nash's relationship deepens, and they work together to uncover the truth about Duncan Hugo's plans. They also face challenges from Tate Dilton, a disgraced police officer, and Lucian Rollins, Nash's friend, who is hiding secrets of his own.
The story is full of twists and turns, including a dramatic confrontation at the library's Book or Treat event, where Lina and her friends are attacked by a group of bikers.

Throughout the story, Lina and Nash's relationship is a central theme, and their chemistry is undeniable. However, their secrets and lies threaten to destroy their trust and ultimately their relationship. The story is a complex exploration of love, trust, and vulnerability, set against the backdrop of a small town with its own secrets and mysteries.

The story is also filled with a cast of colorful characters, including Mrs. Tweedy, a feisty elderly woman who is always up for a fight, and Naomi, a kind-hearted community outreach coordinator who becomes Lina's friend.

The story is a gripping and emotional ride, with a complex plot that keeps readers guessing until the end.

Table 12: Zero-shot summary.

Original HMerge Summary

Lina Solavita, an investigator for Pritzger Insurance, moved to Knockemout, Virginia, where she became entangled in a complicated situation involving Nash Morgan, the police chief. Nash was struggling to come to terms with his new reality after being shot in the line of duty and was feeling disconnected from his life. Lina's presence sparked a series of events, including a confrontation with US Marshal Nolan Graham, who was shadowing Nash. Lina's past was marked by a medical condition and a complicated relationship with physical touch, which made her interactions with Nash tense and awkward.

As Lina settled into her new apartment, she received a call from her mother, Bonnie, who was checking in on her. Lina reassured her that she was fine and that she was enjoying her time in Knockemout. However, Lina's mother was not convinced, and she was already making plans to visit her daughter soon. Lina's conversation with her mother was interrupted by a knock at the door. It was Knox Morgan, Nash's brother, who was there to talk to Nash about something. Nash was not interested in talking, but Knox was insistent. As Knox headed to Nash's apartment, Lina was left to wonder what was going on between the two brothers.

Despite the challenges she was facing, Lina was determined to make the most of her time in Knockemout. She started to feel a connection with Nash, and she was hoping to help him heal and find his way again. As she looked around her new apartment, Lina felt a sense of possibility and promise. She was not sure what the future held, but she was ready to face whatever came her way.

Nash was struggling to come to terms with his past and his feelings for Lina. He was haunted by nightmares and was trying to find a way to heal and move forward. Lina was a strong and independent character, but she was also vulnerable and guarded. Nash was a complex character with a troubled past, but he was also kind and heroic. The chemistry between them was palpable, and their interactions were filled with tension and humor. As Lina navigated her feelings for Nash, she must confront her own vulnerabilities and learn to trust him. Nash's brother, Knox, was a constant presence in the background, offering advice and guidance. However, Nash was determined to forge his own path and make his own decisions. The investigation into the police officer's misconduct and the mysterious henchmen continued to unfold, and it was unclear what the outcome would be. Lina's friends, Stef, Naomi, and Sloane, were trying to support her through her struggles. They were also dealing with their own personal issues, including a recent abduction and a complicated relationship with Nash's brother, Knox. In a separate storyline, Nolan, a marshal, was trying to get Lina to open up about her past. He was a supportive friend and was trying to help her heal and move forward. The two had a non-date drink together, and Lina began to open up about her feelings and experiences.

As the story unfolded, it became clear that Nash and Lina's relationship was complex and multifaceted. They were drawn to each other, but their past experiences and personal issues made it difficult for them to be together. The investigation into the police officer's misconduct and the mysterious henchmen continued to unfold, and it was unclear what the outcome would be. Lina was kidnapped by Cereal Aisle Guy, who was revealed to be Nikos, and taken to a barn-like building where she met Tate Dilton, a dirty cop, and Duncan Hugo, the mastermind. Hugo revealed that Dilton was hired to shoot Nash, but instead shot him in cold blood on the highway. Hugo planned to use Lina to lure Nash to the barn, where he would be killed. Lina tried to stall and gather information, learning that Hugo's goal was to take over his father's business and that Dilton had a personal vendetta against Nash.

Nash and his friends were searching for Lina, following a lead on a partial plate number that Waylay memorized. They were coordinating with the state police and setting up traffic stops to find Lina and the unsub. In a desperate attempt to escape, Lina used a microphone to call for help, reaching out to a group of children playing an online video game. The kids, including a girl named Brecklin, were initially skeptical but eventually believed Lina's plea for help. One of the players, Waylay, recognized Lina's name and realized that her uncle, Nash, was the chief of police. Waylay joined the game and helped Lina by providing crucial information to Nash. Meanwhile, Nash was searching for Lina and had tracked down a lead to a Ford Fusion belonging to Mark Nikos.

Nash, along with his brother Knox and friend Nolan, approached the farm, where they found Lina being held by Dilton and Hugo. A shootout ensued, and Lina managed to escape by using a pressure washer to defend herself. She then stabbed Hugo with a pitchfork and made her way to a Porsche, where she met up with Knox and Nolan. As they tried to escape, Dilton shot Nolan, and Lina and Knox took cover behind a tractor. Nash arrived and proposed to Lina, who accepted. Knox then took Lina and Nolan to safety, while Nash stayed behind to confront Dilton. Nash was able to outmaneuver Dilton and shot him three times, killing him.

The story then cut to a wedding ceremony, where Knox was marrying Naomi. Lina was in attendance, and Nash was serving as a groomsman. The ceremony was emotional, with many guests in tears. After the ceremony, Nash and Lina shared a dance, and he told her that he'd been thinking about their future together. He revealed that he got a tattoo of angel wings on his butt, which Lina found romantic.

As they danced, Nash's father, Duke, approached them and congratulated them on their engagement. Duke was sober and seemed to be making amends for

his past mistakes. Nash and Lina then sneaked away to Nash's old bedroom, where they shared a romantic moment. However, their intimacy was interrupted

by Lina's parents, who had arrived unexpectedly. Nash was caught off guard, but Lina was relieved that her parents had finally met Nash.

Table 13: Original HMerge Summary.

Cite-HMerge Summary

Nash Morgan, a police chief, is struggling to cope with the emotional aftermath of being shot in the line of duty, and his emotional numbness is affecting his relationships with those around him. He meets his new neighbor, Angelina Solavita, also known as Lina, who is an insurance agent and his brother's ex-girlfriend from college, and despite initial reservations, they are drawn to each other.

Lina is investigating a case involving a stolen car, and her search leads her to cross paths with Nash, who is still recovering from his injuries and struggling to come to terms with his feelings about her. As they spend more time together, they grow closer, but their relationship is complicated by Nash's past trauma and Lina's secrets. Lina is searching for a car that Duncan Hugo stole and is looking for any leads in the crime scene files. She is also being pursued by Nash, who is concerned about her safety and thinks she should not be involved in the investigation.

Lina observes a fight between two brothers, Wendell Baker and his brother, and calls 911, but before the police arrive, she is abducted by two people in a van. They take her to meet Grim, the leader of a motorcycle club, who has information about Hugo. Grim reveals that Hugo is still in town and is planning to take over his father's crime business. Lina is frustrated that she is not being included in the plans to deal with Hugo and decides to take matters into her own hands.

She attends a Halloween party at the library, where she runs into an old friend from high school, Angie, and her family. Angie's son, Austin, is recovering from leukemia, and Angie has been thinking about how she can be a better parent to him. Lina and Angie reconnect, and Angie apologizes again for not being a better friend. Lina reflects on how she pushed her friends away after her own health issues and how she is grateful for Angie's apology.

Nash and his friends are preparing for a confrontation with Hugo and his men. Lina is trying to stay one step ahead of Hugo and his plans. She is also dealing with her feelings for Nash and trying to navigate their complicated relationship.

Angie's husband is supportive of her reconnecting with Lina and is grateful for her strength and resilience. Lina and Angie are rebuilding their friendship and supporting each other.

The story is ongoing, with Lina and Nash working together to take down Hugo and his men. Lina is a strong and determined protagonist who is not afraid to take risks. She is also vulnerable and relatable, with a deep emotional life. The story is a thrilling and romantic ride, with many surprises along the way. Lina and Nash's relationship is complicated and passionate, with a deep emotional connection. The story is full of suspense and action, with many twists and turns.

Angelina, dressed as Nancy Drew, is at a Halloween party with her friends Naomi and Sloane, when they are approached by Nash Morgan and his brothers, Knox, Nolan, and Lucian. The group is interrupted by a drunk man who grabs Sloane, leading to a confrontation that ends with Nash arresting the man and his friend, who had also gotten into a fight with Nolan. After the altercation, Nash asks Angelina to dance, and despite initial hesitation, she agrees. As they dance, the tension between them becomes palpable, and Nash eventually pulls her away to a quiet room, where they share a passionate kiss.

Angelina had a life-changing experience with Nash, which left her questioning her feelings and the connection they shared. The night before, Nash had been on his knees, pleasuring her, and she had experienced multiple orgasms, which was a new and overwhelming experience for her. After their intense encounter, Nash was still confident and self-assured, which only added to Angelina's confusion and frustration. Nash asked her what she needed to feel safe with him, and Angelina struggled to articulate her feelings, unsure of how to process the intensity of their connection. Nash and Angelina made a deal to work together and be open with each other. Nash later reflected on their encounter and realized that he had been given a second chance with Angelina. He sent her a playful text message, teasing her about her exhaustion and inviting her to dinner. Angelina responded, and they exchanged banter, with Nash inviting her to dinner and Angelina agreeing. Nash also had a conversation with Sloane, who had discovered that he and Angelina had slept together.

Nash and Angelina's dinner date ended on a romantic note, with Nash and Angelina feeling closer and more connected. Angelina revealed that she wanted to be part of the team working to catch Duncan Hugo and help keep Nash safe. Nash was touched by her concern and willingness to help. He realized that Angelina cared about him deeply and was willing to take risks to protect him.

Table 14: Cite-HMerge Summary.

Extract-Support Summary

In the small town of Knockemout, Police Chief Nash Morgan is struggling to cope with the aftermath of being shot in the line of duty. He's emotionally numb and feeling useless, going through the motions of his job without any real passion or interest. His life is further complicated by the arrival of his brother's ex-girlfriend, Angelina "Lina" Solavita, who moves into the apartment next door.

Lina is a beautiful and charming woman who works as an insurance investigator, and Nash finds himself drawn to her despite his emotional numbness. Lina is in town investigating a case involving a list of law enforcement officers and informants, including Nash, who was targeted by Duncan Hugo, the son of a wealthy and influential man. She's also on the hunt for a stolen 1948 Porsche 356 convertible, which is priceless to her client.

As Lina navigates her investigation, she befriends Naomi, Knox's fiancée, and Sloane, a librarian, who support her and help her through her tough time. Nash, still recovering from his injuries and struggling with memory loss and panic attacks, is haunted by the memory of the shooting and is experiencing flashbacks. Lina is a calm and soothing presence in his life, and he finds himself feeling more at ease around her.

As Nash and Lina navigate their growing attraction, they must also confront the secrets and mysteries that surround them. Lina's investigation leads her to a row home, where she observes a confrontation between two brothers, Wendell Baker and his brother. The situation escalates, and Lina is abducted by a group of people in ski masks, who turn out to be associates of a man named Grim. Grim is a leader of a motorcycle club and has information about Duncan Hugo's whereabouts.

As Nash and Lina's relationship deepens, they grow closer, and Lina begins to confront her fears and consider the possibility of opening up to Nash and her parents. The investigation into Duncan Hugo's whereabouts continues, with the team working together to gather information and stay safe. The team comes up with a plan to use Nash as bait to lure Hugo out of hiding.

Lucian Rollins, a wealthy and influential man, is concerned about Lina's intentions and warns her to stay away from Nash. However, Lina is not intimidated and fires back at Lucian, accusing him of trying to manipulate her and warning him to respect Nash's boundaries. Despite the tension between them, Lucian eventually becomes an ally to Nash and Lina, and he helps them in their quest to bring Hugo to justice.

As the team prepares for the showdown with Hugo, tensions are running high. Knox is worried about the safety of his fiancée, Naomi, and her daughter, Waylay, while Lucian is concerned about the potential risks of the plan. Despite these concerns, the team is determined to bring Hugo to justice.

In the end, Nash and Lina's relationship is put to the test as they work together to take down Hugo and bring him to justice. They face challenges and obstacles, but their connection remains strong. As they navigate the ups and downs of life, they realize that their love is the key to overcoming any obstacle. Nash proposes to Lina in the midst of the chaos, and she accepts.

The story concludes with Nash and Lina emerging from a chaotic and intense situation, surrounded by police and other responders, as a new day begins. Throughout the story, Nash struggles with his emotions after a traumatic event where a woman was severely injured in a car accident. He feels guilty and helpless, remembering the death of his mother in a similar accident. He pushes away his girlfriend, Lina, saying he needs space and doesn't want her around anymore. However, Lina comforts Nash, telling him that he is a hero for facing his demons every day to protect his town and its people.

The story also explores Nash's complicated relationship with his father, Duke, who has been absent for a while but is making an effort to reconnect with his sons. Nash is wary at first, but he's happy to see his father making progress and is grateful for his presence at Knox's wedding.

Table 15: Extract-Support Summary.

Extract-Retrieve Summary

Sloane, a former sophomore, and Lucian, a former senior, were next-door neighbors. Sloane had bought her house from her parents when they moved, while Lucian kept his mom's place. On one occasion, Lou and Amanda volunteered to walk Liza J home, sparking a discussion about who would foster her.

Anthony Hugo, a notorious crime lord, controlled a vast territory that included Washington, DC, and Baltimore. Naomi and Sloane simultaneously expressed their interest in fostering Liza J. Knox and Lucian stood alongside Nash and Nolan, while Officer Dilton revealed that Mr. Peters was Knockemout's solicitor. Lucian proposed a theory that Anthony had helped his son leave the country.

The community had come together to organize a party, with library patrons buying tickets and business sponsors, persuaded by Sloane, donating food and drinks. The library retained the profits from the event. Special Agent Idler had a personal vendetta against Anthony Hugo and was determined to bring him to justice.

Naomi and Knox, along with their extended families, were instrumental in maintaining order in the community. However, Nash Morgan posed a significant threat to women everywhere. Naomi's soon-to-be husband was reminded to take a time-out as Naomi, Sloane, and Mrs. Tweedy approached.

Tallulah and Justice St. John, along with Gael and his husband, Isaac, occupied a table, enjoying their monthly double date. They were seated next to Fi and her husband. Naomi and Sloane were taking their assignments seriously, and Sloane raised her hand to contribute to the discussion. According to the Knockemout grapevine, Knox, Nolan, Lucian, and others had gathered in Knox's secret lair office. Knox's gaze met his brother's, and they shared a long, meaningful look. Jeremy Trent, the former captain of the baseball team, had beaten out Dilton for homecoming king during their senior year of high school.

Knox drove Waylay and Liza J home to Naomi, who was tracked as being at home. Naomi and Sloane also approached with a proposal, inquiring if they would be interested in helping with their new venture. Naomi's wedding was scheduled for the next day, and Waylay was just a child. Lina was spotted in the photo booth again with Sloane and Fi, as they celebrated the upcoming wedding. The atmosphere was filled with discussions of real estate, marriage, and babies.

Liza J's presence was a reminder of the challenges that lay ahead. Lucian and a private security team were tasked with guarding them, ensuring their safety. Lucian was seen clutching his prized Smith & Wesson six-shooter, a symbol of his commitment to protecting those he cared about. He nodded and followed Melissa inside, as they worked together to address the looming threat. Duncan Hugo was wanted for attempted murder of a law enforcement officer, domestic violence, and assault.

Concerns about Naomi's safety were raised, and Lucian reassured everyone that she was being protected. Hugo's careless attitude was a stark contrast to the gravity of the situation. A warning was issued to call Nash and inform him that Duncan Hugo was sending Tate Dilton to their house. The situation was dire, and it was essential to take immediate action. The challenge of explaining the situation to a twelve-year-old was daunting. Duncan Hugo was in custody, but the threat to their safety remained.

A plan was devised to fly in Liza J's family to surprise her, a gesture that would bring some joy amidst the chaos. The family tradition of grand gestures was evident, as Lou recalled surprising Mandy with a three-week cruise. Knox, the groom, was the focus of attention, and it was essential to ensure his safety, as well as that of Nolan and Liza J. Jeremiah and Lou asked in unison if they should take them to a safe location.

The atmosphere was tense, with the sound of a Ford Fusion, belonging to Mark Nikos, a reminder of the outside world. As the events unfolded, it became clear that the community was coming together to support each other. Naomi and Sloane's commitment to their assignments was evident, and their dedication to the community was inspiring. The presence of Lucian and the private security team provided a sense of reassurance, as they worked together to address the threats posed by Duncan Hugo and Nash Morgan. The celebration of Naomi's wedding was a beacon of hope, a reminder that even in the face of adversity, the community could come together and find joy.

Table 16: Extract-Retrieve Summary.

Lina Solavita is an insurance investigator who moves to Knockemout, a small town in northern Virginia.	✓
Lina is in Knockemout to find a 1948 Porsche 356 convertible that was stolen from a client.	✓
Lina is staying in an apartment next to Nash Morgan, who is the chief of police in Knockemout.	✓
Nash Morgan is recovering from a shooting that left him with bullet wounds and memory loss.	✓
Nash Morgan struggles to cope with his new reality and is haunted by the sound associated with the moment he was shot.	✓
Lina is initially attracted to Nash Morgan, but their relationship is complicated by his vulnerability and her job.	✗
Lina's job as an insurance investigator requires her to keep secrets.	—
Lina is dealing with her own emotional baggage, including a heart condition that almost killed her at the age of 15.	✓
A U.S. marshal named Nolan Graham is watching Nash Morgan due to a threat against him.	✓
Lina is caught in the middle of the situation between Nash Morgan and Nolan Graham.	✓
As Lina and Nash spend time together, they develop a strong connection.	✓
Lina and Nash's differences and secrets threaten to tear their relationship apart.	✗
Lina's past, including her history with Nolan Graham and her job, is revealed gradually throughout the story.	✓
Nash Morgan feels torn between his growing feelings for Lina and his need for honesty and trust.	—
Lina's investigation leads her into danger, prompting Nash to confront his vulnerabilities.	—
Lina and Nash are eventually kidnapped by a group of bikers working for Grim, a motorcycle club leader.	✗
Grim reveals that Duncan Hugo, the man who shot Nash, is still in town and planning to take over his father's criminal empire.	✓
The relationship between Lina and Nash deepens as they work together to uncover Duncan Hugo's plans.	✓
Lina and Nash face challenges from Tate Dilton, a disgraced police officer, and Lucian Rollins, who is Nash's friend.	✓
The story includes a dramatic confrontation at a library's Book or Treat event where Lina and her friends are attacked by bikers.	✗
Lina and Nash's relationship is a central theme of the story, marked by undeniable chemistry.	✓
Secrets and lies threaten to destroy the trust between Lina and Nash, potentially ending their relationship.	✗
The story is a complex exploration of love, trust, and vulnerability set in a small town with its own secrets.	—
The cast of characters includes Mrs. Tweedy, a feisty elderly woman, and Naomi, a kind-hearted community outreach coordinator who becomes Lina's friend.	✓
The narrative is described as a gripping and emotional ride with a complex plot that keeps readers guessing until the end.	—

Table 17: Fact annotations for the Zero-shot summary.

Lina Solavita is an investigator for Pritzger Insurance, and she moved to Knockemout, Virginia.	✓
Nash Morgan is the police chief of Knockemout, Virginia, who was shot in the line of duty and was struggling to accept his new reality.	✓
Lina's presence in Knockemout sparked a confrontation with US Marshal Nolan Graham, who was shadowing Nash.	✗
Lina has a past marked by a medical condition and a complicated relationship with physical touch, which made her interactions with Nash awkward.	✗
Lina received a call from her mother, Bonnie, who was checking in on her while she was settling into her new apartment.	✓
Lina reassured her mother that she was fine and was enjoying her time in Knockemout, but Bonnie was not convinced and planned to visit soon.	✓
Knox Morgan, Nash's brother, knocked on Lina's door to talk to Nash about a matter that he found important.	✗
Knox insisted on discussing something with Nash, even though Nash was not interested in talking.	—
As Lina waited for Nash and Knox's conversation to conclude, she was left wondering about the relationship between the two brothers.	—
Despite her challenges, Lina was determined to make the most of her time in Knockemout.	✓
Lina started to feel a connection with Nash and hoped to help him heal and find his way again.	—
Nash was struggling with his feelings for Lina while haunted by nightmares and trying to heal from his past.	✗
Lina is portrayed as a strong yet vulnerable and guarded character in the story.	✓
Nash is depicted as a complex character with a troubled past who is also kind and heroic.	✓
The chemistry between Lina and Nash is described as palpable, with their interactions filled with tension and humor.	—
As Lina navigated her feelings for Nash, she confronted her own vulnerabilities and learned to trust him.	✓
Nash's brother, Knox, was a constant presence in the background, providing advice and guidance to him.	✓
Nash was determined to make his own decisions and forge his own path despite his brother's influence.	—
The investigation into police officer misconduct and the activities of mysterious henchmen was ongoing in Knockemout.	✓
Lina's friends, Stef, Naomi, and Sloane, were supporting her while dealing with their own personal issues, including a recent abduction.	✗
Nolan Graham, a US Marshal, was trying to get Lina to open up about her past and was acting as a supportive friend.	✗
Lina and Nolan had a non-date drink together, during which Lina began to open up about her feelings and experiences.	✗
The relationship between Nash and Lina was complex and multifaceted, shaped by their past experiences.	✓
Lina was kidnapped by someone known as Cereal Aisle Guy, who was revealed to be Nikos.	✓
Lina was taken to a barn-like building where she met Tate Dilton, a corrupt cop, and Duncan Hugo, the mastermind behind the kidnapping.	✓
Duncan Hugo revealed that he hired Tate Dilton to shoot Nash, but Dilton shot him in cold blood on the highway instead.	✓
Duncan's plan was to use Lena to lure Nash to the barn, where Nash would be killed.	✗
While held captive, Lina learned about Duncan's goal to take over his father's business and Tate's personal vendetta against Nash.	✓
Nash and his friends were searching for Lina by following a lead on a partial plate number that Waylay memorized.	✓
Waylay, a character in the story, helped Lina by recognizing her name and informed Nash of her situation during an online game.	✗
Nash tracked down a lead to a Ford Fusion belonging to Mark Nikos, as he searched for Lina.	✓
Nash, along with Knox and Nolan, approached the farm where Lina was being held by Tate and Duncan.	✓
A shootout occurred at the farm during the rescue attempt, where Lina managed to use a pressure washer to defend herself.	✗
Lina stabbed Duncan Hugo with a pitchfork and escaped to a Porsche, where she met up with Knox and Nolan.	✗
During the escape attempt, Dilton shot Nolan, and Lina and Knox sought cover behind a tractor.	✓
Nash proposed to Lina, and she accepted his proposal amid the ongoing chaos.	✓
Knox managed to take Lina and Nolan to safety while Nash stayed behind to confront Tate Dilton.	✓
Nash successfully killed Tate Dilton by shooting him three times after outmaneuvering him.	✓
The story features a wedding ceremony where Knox marries Naomi, with Lina in attendance and Nash serving as a groomsman.	✗
After the wedding ceremony, Nash and Lina shared a dance, during which he expressed thoughts about their future together.	✓
Nash revealed to Lina that he got a tattoo of angel wings on his butt, which she found romantic.	✓
During their dance, Nash's father, Duke, congratulated them on their engagement and seemed to be making amends for past mistakes.	✓
Nash and Lina sneaked away to Nash's old bedroom, where they shared a romantic moment that was interrupted by the unexpected arrival of Lina's parents.	✓
Lina felt relieved that her parents had finally met Nash, even though he was caught off guard by their unexpected presence.	—

Table 18: Fact annotations for the original HMerge summary.

Nash Morgan is a police chief who is struggling to cope with the emotional aftermath of being shot in the line of duty. His emotional numbness is affecting his relationships with those around him.	✓
Angelina Solavita, also known as Lina, is Nash's new neighbor and an insurance agent. Lina is also Nash's brother's ex-girlfriend from college.	✗
Lina is investigating a case involving a stolen car, which leads her to cross paths with Nash. Nash is still recovering from his injuries during their encounters.	✓
As Nash and Lina spend more time together, they grow closer, but their relationship is complicated by Nash's past trauma and Lina's secrets. Nash's emotional state is affected by his experiences as a police officer.	✓
Lina is searching for a car that Duncan Hugo stole and is looking for leads in the crime scene files. She feels frustrated that she is not being included in the plans to deal with Hugo.	✗
Lina observes a fight between two brothers, Wendell Baker and his brother, and calls 911. Before the police arrive, she is abducted by two people in a van.	✓
The two abductors take Lina to meet Grim, the leader of a motorcycle club. Grim has information about Duncan Hugo, who is still in town and planning to take over his father's crime business.	✓
Nash is concerned about Lina's safety and believes she should not be involved in the investigation regarding Hugo. He is pursuing her to ensure her well-being.	✓
Lina attends a Halloween party at the library and runs into an old friend from high school, Angie. Angie is coping with her son Austin's recovery from leukemia and reflects on how to be a better parent.	✓
Lina reconnects with Angie, who apologizes for not being a better friend. Lina reflects on how she pushed her friends away after her own health issues.	✓
Nash and his friends are preparing for a confrontation with Duncan Hugo and his men. Lina is trying to stay one step ahead of Hugo's plans.	—
Lina and Nash's relationship is complicated by her feelings and past experiences. She is dealing with her vulnerabilities while navigating their connection.	✓
Angelina, dressed as Nancy Drew, is approached at the Halloween party by Nash Morgan and his brothers, Knox, Nolan, and Lucian. Their group is interrupted by a drunk man grabbing Sloane.	✗
Nash arrests the drunk man and his friend, who had also fought with Nolan. After the altercation, Nash asks Angelina to dance, which she initially hesitates to do.	✗
While dancing, the tension between Nash and Angelina becomes palpable, leading Nash to pull her away to a quiet room for a passionate kiss.	—
Angelina reflects on her life-changing experience with Nash, which leaves her questioning her feelings. The night before, they had an intimate encounter that overwhelmed her.	✓
Nash and Angelina made a deal to work together and be open with each other after their intense encounter. Nash realizes he has been given a second chance with Angelina.	✓
Nash sends Angelina a playful text message after their encounter, teasing her about her exhaustion. They banter and agree to go out to dinner together.	—
Nash speaks with Sloane, who discovers that he and Angelina had slept together. Their dinner date ends on a romantic note, enhancing the connection between them.	—
Angelina expresses her desire to help catch Duncan Hugo and keep Nash safe. Nash is touched by her concern and realizes that Angelina cares about him deeply.	✓

Table 19: Fact annotations for the Cite-HMerge summary.

Police Chief Nash Morgan struggles to cope with the aftermath of being shot in the line of duty in the small town of Knockemout.	✓
Nash is emotionally numb and feeling useless while going through the motions of his job without passion or interest.	✓
Angelina "Lina" Solavita, Nash's brother's ex-girlfriend, moves into the apartment next door to Nash.	✓
Lina is a beautiful and charming woman who works as an insurance investigator.	✓
Nash is drawn to Lina despite his emotional numbness.	✓
Lina is in Knockemout investigating a case involving law enforcement officers and informants, including Nash, who was targeted by Duncan Hugo.	✗
Duncan Hugo is the son of a wealthy and influential man who presents a threat to Nash and others.	✓
Lina's investigation includes searching for a stolen 1948 Porsche 356 convertible that is priceless to her client.	✓
Lina befriends Naomi, Knox's fiancée, and Sloane, a librarian, while navigating her investigation.	✓
Nash is recovering from his injuries and is experiencing memory loss and panic attacks.	✓
Nash is haunted by flashbacks of the shooting incident.	✓
Lina acts as a calm and soothing presence in Nash's life, helping him feel more at ease.	—
Nash and Lina confront secrets and mysteries surrounding them as their attraction grows.	✓
Lina observes a confrontation between Wendell Baker and his brother at a row home during her investigation.	✓
Lina is abducted by a group of people in ski masks during the confrontation, who turn out to be associates of a man named Grim.	✓
Grim is the leader of a motorcycle club who possesses information about Duncan Hugo's whereabouts.	✓
As Nash and Lina's relationship deepens, Lina begins to confront her fears about opening up to Nash and her parents.	—
The investigation into Duncan Hugo's whereabouts continues as a team works together to gather information and ensure their safety.	✓
The team plans to use Nash as bait to lure Duncan Hugo out of hiding.	✗
Lucian Rollins, a wealthy and influential man, expresses concern over Lina's intentions and warns her to stay away from Nash.	✗
Lina confronts Lucian for trying to manipulate her and warns him to respect Nash's boundaries.	✓
Despite initial tension, Lucian becomes an ally to Nash and Lina, assisting them in their quest to bring Hugo to justice.	✓
As the team prepares for a showdown with Duncan Hugo, tensions rise regarding the safety of Naomi and her daughter, Waylay.	—
The team is determined to bring Duncan Hugo to justice despite their concerns.	✓
Nash and Lina's relationship faces tests as they work together to confront Duncan Hugo.	✓
Nash proposes to Lina amidst chaos, and she accepts his proposal.	✓
The story concludes with Nash and Lina surrounded by police and responders, emerging from an intense situation as a new day begins.	✗
Nash struggles with feelings of guilt and helplessness after a traumatic incident where a woman was severely injured in a car accident.	✓
The memory of his mother's death in a similar accident adds to Nash's guilt.	✗
Nash pushes away Lina, telling her he needs space and does not want her around anymore.	✓
Despite Nash's rejection, Lina comforts him, reminding him that he is a hero for facing his demons.	✓
The story explores Nash's complicated relationship with his father, Duke, who has been absent but is trying to reconnect with his sons.	✓
Nash feels wary but ultimately happy to see his father making progress and is grateful for his presence at Knox's wedding.	✗

Table 20: Fact annotations for the Extract-Support summary.

Sloane is a former sophomore who bought her house from her parents when they moved.	✓
Lucian is a former senior who kept his mom's house.	✓
Lou and Amanda volunteered to walk Liza J home, which sparked a discussion about who would foster her.	—
Anthony Hugo is a notorious crime lord who controls a vast territory that includes Washington, DC, and Baltimore.	✓
Naomi and Sloane simultaneously expressed interest in fostering Liza J.	—
Knox and Lucian stood alongside Nash and Nolan while Officer Dilton revealed that Mr. Peters was Knockemout's solicitor.	×
Lucian proposed a theory that Anthony Hugo had helped his son leave the country.	✓
The community organized a party with library patrons buying tickets and business sponsors donating food and drinks, persuaded by Sloane.	✓
The library retained the profits from the event.	✓
Special Agent Idler has a personal vendetta against Anthony Hugo and is determined to bring him to justice.	✓
Naomi and Knox, along with their extended families, helped maintain order in the community.	✓
Nash Morgan posed a significant threat to women everywhere.	×
Naomi was reminded to take a time-out as she, Sloane, and Mrs. Tweedy approached her soon-to-be husband.	—
Tallulah and Justice St. John, along with Gael and his husband Isaac, occupied a table enjoying their monthly double date.	✓
Naomi and Sloane were taking their assignments seriously, and Sloane raised her hand to contribute to the discussion.	—
Knox, Nolan, Lucian, and others gathered in Knox's secret lair office according to the Knockemout grapevine.	×
Knox shared a long, meaningful look with his brother.	×
Jeremy Trent, the former captain of the baseball team, beat Dilton for homecoming king during their senior year.	✓
Knox drove Waylay and Liza J home to Naomi, who was tracked as being at home.	✓
Naomi and Sloane approached others with a proposal about their new venture.	—
Naomi's wedding was scheduled for the next day while Waylay was just a child.	✓
Lina was spotted in the photo booth with Sloane and Fi as they celebrated the upcoming wedding.	✓
Discussions about real estate, marriage, and babies filled the atmosphere.	×
Liza J's presence reminded the community of the challenges that lay ahead.	×
Lucian and a private security team were tasked with ensuring the safety of Naomi, Sloane, and others.	✓
Lucian was seen clutching his prized Smith & Wesson six-shooter, symbolizing his commitment to protection.	×
Lucian nodded and followed Melissa inside to address the looming threat.	×
Duncan Hugo was wanted for attempted murder of a law enforcement officer, domestic violence, and assault.	✓
Concerns about Naomi's safety were raised, and Lucian reassured everyone that she was being protected.	—
Duncan Hugo's careless attitude contrasted sharply with the gravity of the situation.	—
A warning was issued to call Nash and inform him that Duncan Hugo was sending Tate Dilton to their house.	—
The situation was deemed dire, necessitating immediate action.	✓
Explaining the situation to a twelve-year-old was seen as daunting.	✓
Duncan Hugo was in custody, but the threat to safety remained.	✓
A plan was devised to fly in Liza J's family to surprise her as a gesture of joy amidst chaos.	—
Lou recalled a family tradition of grand gestures, like surprising Mandy with a three-week cruise.	✓
Knox, the groom, was the focus of attention, and it was essential to ensure his safety along with Nolan and Liza J.	—
Jeremiah and Lou asked unanimously if they should take Knox and others to a safe location.	—
The atmosphere was tense with a sound of a Ford Fusion belonging to Mark Nikos, reminding everyone of the outside world.	—
Community support was evident as events unfolded.	✓
Naomi and Sloane's commitment to their assignments showed dedication to the community.	✓
The presence of Lucian and the private security team provided reassurance against threats posed by Duncan Hugo and Nash Morgan.	×
The celebration of Naomi's wedding symbolized hope and community unity in the face of adversity.	×

Table 21: Fact annotations for the Extract-Replace summary.