Distilling Opinions at Scale: Incremental Opinion Summarization using XL-OPSUMM

Anonymous ACL submission

Abstract

Opinion summarization in e-commerce encapsulates the collective views of numerous users about a product based on their reviews. Typically, a product on an e-commerce platform has thousands of reviews, each review comprising around 10-15 words. While Large Language Models (LLMs) have shown proficiency in summarization tasks, they struggle to handle such a large volume of reviews due to context limitations. To address this, we propose a scalable framework called **XL-OPSUMM** that generates summaries incrementally with the help of ASPECT DICTIONARY (Refer to Section 3). However, the existing test set, AMASUM has only 560 reviews per product on average. Due to the lack of a test set with thousands of reviews, we created a new test set called **XL-FLIPKART** by gathering data from the Flipkart website and generating summaries using GPT-4[1]. Through various automatic evaluations and extensive analysis, we evaluated the framework's efficiency on two datasets, AMASUM and XL-FLIPKART. Experimental results show that our framework, XL-OPSUMM powered by LLAMA-3-8B-8K, achieves an average ROUGE-1 F1 gain of $\mathbf{4.38}\%$ and a ROUGE-L F1 gain of $\mathbf{3.70}\%$ over the next best-performing model.

1 Introduction

E-commerce websites are valuable sources of product reviews, aiding users in well-informed purchasing decisions. Yet, sifting through numerous reviews can be daunting and time-consuming. Opinion summarization offers a solution by summarizing the opinions presented in product reviews (Hu and Liu, 2006; Wang and Ling, 2016; Angelidis and Lapata, 2018; Siledar et al., 2023). However, their utility is limited when confronted with the vast number of reviews, typical of e-commerce platforms. Recent advancements in opinion summarization (Bhaskar et al., 2023; Hosking et al., 2023) address this by scaling systems to accommodate a larger number of reviews, yet they still fall short of fully harnessing the vast array of reviews often numbering in the thousands.

Recent studies have demonstrated that Large Language Models (LLMs) can generate effective opinion summaries in zero-shot prompt settings (Siledar et al., 2024a). However, when dealing with large contexts, LLMs often struggle to retrieve relevant information from the middle of the context (Liu et al., 2023). Furthermore, despite their ability to process a large number of tokens, LLMs are constrained by context limits and cannot accommodate the entire set of reviews, which typically number in the thousands.

To address this issue, incremental and hierarchical approaches have been proposed by Chang et al. (2023). Nonetheless, these methods may not effectively manage conflicting opinions about specific aspects across different chunks of reviews while updating the summary.

The unavailability of any large-scale (ranging in thousands of reviews) test sets hinders progress in this area. To address these issues, we first create XL-FLIPKART, a test set containing $\sim$ 3680 reviews on average per product for 25 products from the Flipkart Website[2]. We employ GPT-4 to annotate summaries (Gilardi et al., 2023; Huang et al., 2023; Siledar et al., 2024b). Next, we propose using an incremental approach to summarize reviews and generate summaries. This we claim has two benefits: (*a*) in the presence of a fresh set of reviews, after a certain period of time (usually the case in the e-commerce domain), our approach emerges as an efficient way of updating summaries,

[1] GPT-4: openai/gpt-4

[2] Flipkart: flipkart.com

and (*b*) does not face context-limit issues which is usually the case when handling such large amount of reviews.

Our contributions are:

1. **XL-FLIPKART**, a large-scale ($\sim$ **3600** reviews on average per product) test set of 25 products gathered from the Flipkart website annotated using GPT-4 (Section 5). *To the best of our knowledge*, this is the first large-scale opinion summarization test set.

2. **XL-OPSUMM**, a large-scale opinion summarization framework that uses an incremental approach capable of generating summaries efficiently using thousands of reviews without any context limitation (Figure 1, Section 3). Experimental demonstrations indicate that our XL-OPSUMM framework powered by LLAMA-3-8B-8K, achieves an average ROUGE-1 F1 gain of **4.38**% and a ROUGE-L F1 gain of **3.70**% over the next best-performing model (Table 3).

3. Qualitative and comparative analysis indicating the efficacy of our XL-OPSUMM framework in handling thousands of reviews for generating comprehensive opinion summaries compared to existing approaches (Sections 7.3 & 7.4).

2 Related Work

Opinion Summarization employs two main approaches: extractive and abstractive. Extractive methods involve selecting the most pertinent sentences directly from the input text, while abstractive techniques generate a condensed version of the opinions expressed.

A Widely used extractive method is the centroid approach, which ranks sentences by relevance to the input text. Another technique is clustering, where sentences are grouped by themes and representative ones are chosen from each cluster. Centroid-based methods include (Radev et al., 2004; Rossiello et al., 2017; Gholipour Ghalandari, 2017), which prioritize sentence selection based on their centrality to the input, and graph-based methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Zheng and Lapata, 2019), which construct graphical representations of the text and extract sentences located at central nodes.

Abstractive opinion summarization is often performed in a self-supervised manner by treating a single review as a pseudo-summary. Various approaches exist for selecting pseudo-summaries and their corresponding input reviews. Bražinskas et al. (2020) employed a random selection of N reviews per entity to construct N pseudo-summary, review pairs. Amplayo and Lapata (2020) sampled a review randomly and generated noisy versions of it as input reviews. Amplayo et al. (2020) used aspect and sentiment distributions to guide pseudo-summary sampling. Elsahar et al. (2021) selected input reviews with high TF-IDF cosine similarity to a randomly sampled pseudo-summary. Wang and Wan (2021) focused on reducing opinion redundancy by learning aspects and sentiment embeddings to generate highly relevant review-pseudo-summary pairs. Im et al. (2021) used a synthetic dataset creation strategy similar to Bražinskas et al. (2020), extending it to multimodal data. Ke et al. (2022) emphasized consistency of aspects and sentiment between reviews and pseudo-summary by using constrained sampling. Finally, Siledar et al. (2023) leveraged lexical and semantic similarities for creating synthetic datasets and Siledar et al. (2024b) uses additional information sources such as product description and question answers of a product to create the synthetic dataset. However, these methods fail to accommodate a substantial volume of review sets as they typically rely on a limited number of input reviews (e.g., 10 reviews) to produce the opinion summary.

Large Scale Opinion Summarization Recent opinion summarization systems such as (Bhaskar et al., 2023; Hosking et al., 2023; Jiang et al., 2023) include a large number of reviews. Bhaskar et al. (2023) explores prompting by testing (OpenAI, 2023) and introduces various pipelines whereas Jiang et al. (2023) introduced a review sampling strategy that uses sentiment analysis and two-stage training scheme to generate the opinion summary. Hosking et al. (2023) encodes the reviews into discrete latent space and then generates the summary by decoding the frequent encodings.

Incremental Summarization (Chowdhury et al., 2024) proposes CoverSumm an algorithm to perform centroid-based extractive opinion

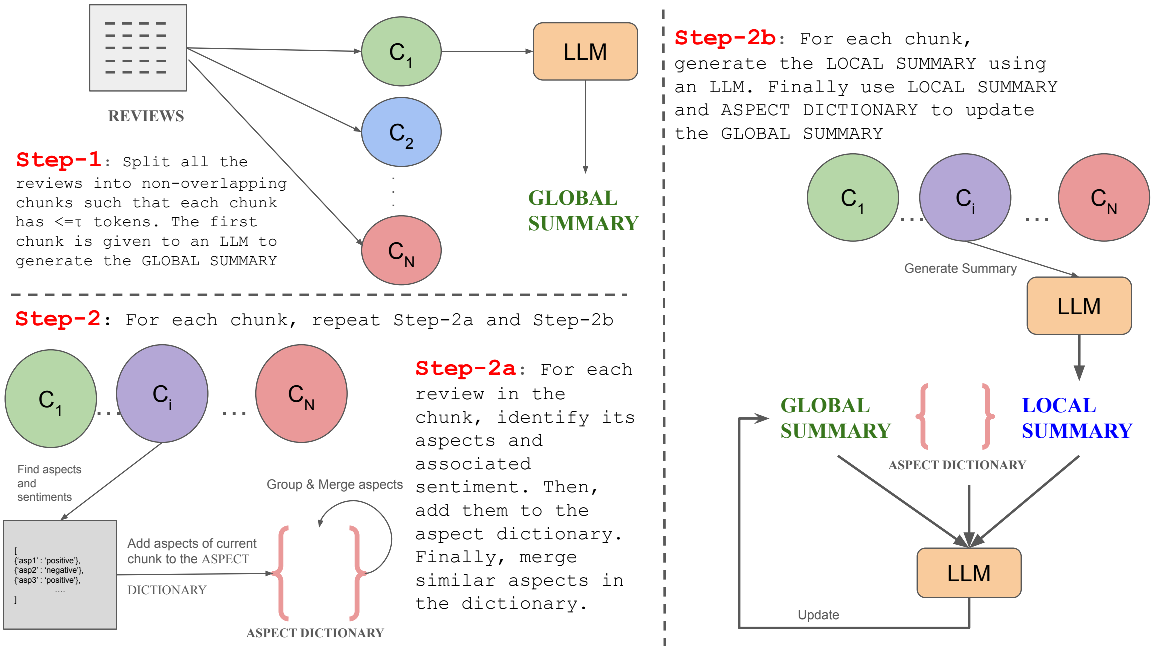

Figure 1: Illustration of our **XL-OPSUMM** framework. First, reviews are divided into non-overlapping chunks based on threshold. Then for each chunk, the ASPECT DICTIONARY is updated, the LOCAL SUMMARY is generated and the GLOBAL SUMMARY is updated as shown above. Refer to the section 3 for more details about this framework

summarization incrementally. (Chang et al., 2023) uses incremental and hierarchical approaches to summarise book-length text. We propose XL-OPSUMM framework for a large-scale opinion summarization system that generates the opinion summary incrementally.

3 XL-OpSumm Framework

To summarize reviews of a product, we split them into non-overlapping chunks, each with up to τ tokens. We then analyze each chunk using three elements: **LOCAL SUMMARY**, **GLOBAL SUMMARY**, and **ASPECT DICTIONARY**. The LOCAL SUMMARY is the summary of all reviews in the current chunk, while the GLOBAL SUMMARY is the summary of all previous chunks. The ASPECT DICTIONARY contains aspects and their corresponding positive, negative, and neutral sentiment counts expressed by users from previous segments. Here are the steps as shown in figure 1 to obtain the final summary for the product:

Step-1: The GLOBAL SUMMARY is initialized with a summary generated by an LLM using all reviews from the first chunk.

Step-2: For each chunk, we repeat the following procedure:

Step-2a: For each review in the chunk, we identify its aspects and corresponding sentiments using the Aspect-Based Sentiment Analyser (ABSA) Model. We then update the ASPECT DICTIONARY by adding the sentiments of aspects in the current chunk to the ASPECT DICTIONARY. To avoid redundancy, we merge similar aspects into a single aspect by encoding the aspect names in the dictionary using Sentence Transformer (Reimers and Gurevych, 2019) and clustering them using the fast clustering algorithm. The sentiment counts of all aspects in one cluster are added together and finally represented using one aspect name.

Step-2b: We use an LLM to generate the summary of the current chunk and assign it to the **LOCAL SUMMARY**. Finally, we use the ASPECT DICTIONARY and the LOCAL SUMMARY to update the GLOBAL SUMMARY. The procedure to update the GLOBAL SUMMARY is presented in appendix A.

After processing all the chunks, the summary in the GLOBAL SUMMARY element is considered as the final summary for the product.

4 Dataset Details

AMASUM (Bražinskas et al., 2021) involves the summarization of reviews of various products from

Previous GLOBAL SUMMARY	LOCAL SUMMARY	Updated GLOBAL SUMMARY
The provided text appears to be a collection of customer reviews for the Realme 8 smartphone. Customers have provided a mix of positive and negative feedback on various aspects such as display, fingerprint sensor, camera quality, battery life, performance, and charging speed. Some users have expressed dissatisfaction with the camera quality and overall performance, while others praised the phone for its display, fingerprint sensor, battery backup, and value for money. It's evident that while the Realme 8 has received some positive feedback, there are also concerns that potential buyers should consider.	Realme 8 smartphone offers sturdy and strong build quality, although its back is prone to fingerprints, necessitating the use of a back case. The rear camera, powered by Sony IMX sensors, delivers excellent results, earning a 5/5 rating. However, the front camera captures only decent pictures. The phone's performance is commendable, thanks to the MTK G 95 processor, which smoothly handles day-to-day applications and gaming. It's well-suited for games like COD, BGMI, and Fortnite. Additionally, the impressive 5,000 mAh battery can easily last up to a day with normal usage, and it supports 30 Watts fast charging, allowing for a full charge within an hour. Lastly, the SUPER AMOLED display enhances the viewing experience, surpassing other LCD or IPS LED displays in the same price range. Overall, the Realme 8 is a reliable choice for those seeking a value-for-money smartphone in this category.	This review highlights the Realme 8 smartphone's performance, camera quality, battery life, and display. The user praises the phone's sturdy build, powerful back camera, seamless performance for gaming, and impressive battery life. They also appreciate the SUPER AMOLED display, which outshines other IPS LED displays in the same price range. However, they note that the front camera's quality is average, and they experience a slight battery degradation after three days. Despite these minor drawbacks, the overall sentiment is positive, and the user recommends the phone based on its value for money and great features. Overall, the Realme 8 receives a rating of 4.5 out of 5 stars.

Table 1: Qualitative Analysis. This table demonstrates how our framework updates summaries. The LOCAL SUMMARY represents the summary of the current chunk, while the Previous GLOBAL SUMMARY encapsulates the summaries of all previously processed chunks. The Updated GLOBAL SUMMARY combines the summaries of all chunks up to and including the current chunk. Conflicting aspect opinions between the Local and Global summaries are shown in red, and new aspects are highlighted in blue. Updated information using the Aspect Dictionary is marked in yellow. For more details, refer to Section 7.3.

	AMASUM	XL-FLIPKART
Average #reviews per entity	560.43	3682.88
Average #sentences per review	3.64	1.63
Average #words per sentence	13.72	10.23

Table 2: Dataset statistics of AMASUM, XL-FLIPKART

Amazon website[3], averaging over 560 reviews per product. In the original dataset, references are categorized into 'verdict', 'pros', and 'cons'. Following Hosking et al. (2023), we merge them to form unified summaries. We then narrowed down the original dataset to four prevalent categories (Electronics, Shoes, Sports & Outdoors, Home & Kitchen) and sampled a subset of 50 entities, resulting in a total of 200 products. Various statistics of the test set are recorded in Table 2.

5 Testset Creation: XL-FLIPKART

The existing AMASUM test set contains approximately 560 reviews per product. However, in a real e-commerce environment, the number of reviews per product typically reaches into the thousands, which is not represented by the AMASUM dataset. To evaluate our XL-OPSUMM framework in a context closer to real-world scenarios, we collected reviews of 25 mobile products from the Flipkart website. As shown in Table 2, each product in this dataset has around $3,680$ reviews on average. This number is nearly 6.5 times greater than the average number of reviews per product in the AMASUM dataset.

Generating summaries for such a large volume of reviews is not only time-consuming but also very challenging for humans. Based on studies by Siledar et al. (2024b) which indicate that humans prefer GPT-generated summaries over those written by humans, we utilized GPT-4-turbo to generate the summaries for the products we collected. The prompt used for generating these summaries with GPT-4-turbo is provided below.

Prompt: `Following are the reviews for a product. Generate a summary of the opinions as a review itself with a word limit of under 100 words. Use information from the given reviews only to generate the summary.`
`reviews: [r1,...,rk]`

[3]Amazon: amazon.in

6 Experiments

6.1 Baseline Models

We evaluate our framework against various baselines, including both abstractive and extractive systems. Important recent state-of-the-art work is mentioned in this section. Refer to the Appendix C for all the other baselines we considered for this work.

6.1.1 Non-LLM Baselines

We evaluated our framework against the following Non-LLM Models

HERCULES$_{\text{EXT}}$ (Hosking et al., 2023) computes extractive summaries by calculating te centroid from each evidence set generated by using HERCULES based on ROUGE-2 F1 score.

BiMeanVAE and COOP (Iso et al., 2022) work by encoding entire reviews into continuous latent vectors. BiMeanVAE takes the average of these encodings while COOP calculates the optimized combination of review encodings.

HERCULES$_{\text{ABS}}$ (Hosking et al., 2023) represents a method that aggregates reviews into summaries by identifying frequent opinions in discrete latent space.

6.1.2 LLM Baselines

We evaluated our framework against the following LLM Models

LLAMA-3-8B-8K[4] is an open source large language model with 8B parameters and 8k context limit.

PHI-3-MINI-3.8B-4K (Abdin et al., 2024) is an open source 3.8B parameter model with 4k context limit

PHI-3-MINI-3.8B-128K (Abdin et al., 2024) is an open source 3.8B parameter model with 128k context limit

LLAMA-3-8B-8K-INCREMENTAL is a method to update the existing summary incrementally using a chunk of reviews (Chang et al., 2023) with the help of LLAMA-3-8B-8K model.

LLAMA-3-8B-8K-HIERARCHICAL is a method of summarizing chunks of reviews and then hierarchically merging the summaries until one summary (Chang et al., 2023) using the LLAMA-3-8B-8K model.

PHI-3-MINI-3.8B-128K-INCREMENTAL is a method to update the existing summary incrementally using a chunk of reviews (Chang et al., 2023) with the help of PHI-3-MINI-3.8B-128K model.

PHI-3-MINI-3.8B-128K-HIERARCHICAL is a method of summarizing chunks of reviews and then hierarchically merging the summaries until one summary (Chang et al., 2023) using the PHI-3-MINI-3.8B-128K model.

6.2 Implementation Details

We conducted all experiments using Nvidia DGX A100 GPUs with 80GB of memory. For the large language models (LLMs) used in our experiments, we set the temperature to 0.8. To populate the aspect dictionary, we employed the Instruct ABSA model (Varia et al., 2023) as our aspect-based sentiment analyzer. Within our framework, we experimented with two LLM options: LLAMA-3-8B-8K and PHI-3-MINI-3.8B-4K (Abdin et al., 2024). When using LLAMA-3-8B-8K, we set the τ value to 4000, whereas for PHI-3-MINI-3.8B-4K, the τ value was scaled down to 2700 due to its context limitation.

7 Results and Analysis

In this section, we show results on various automatic reference-based metrics and reference-free metrics as well. We also analyze our model's performance qualitatively and comparatively with other models' summaries.

7.1 Automatic Evaluation

The evaluation of the generated summaries is conducted using the ROUGE-1,2,L F1 score (R1, R2 & RL)(Lin, 2004) and BERT-F1 score(Zhang et al., 2019). Refer to Appendix B for more description of these metrics. It is noted that there is a possibility that the LLAMA-3-8B-8K and PHI-3-MINI-3.8B-4K models may not be able to handle the input tokens in XL-FLIPKART and AMASUM datasets. To address this, the input was truncated, and the maximum number of tokens that the models could handle was used to obtain the results.

[4] Llama-3:meta/llama-3

	abs?	**Model**	**AmaSum** R1 ↑	R2 ↑	RL ↑	BERT-F1 ↑	**XL-FLIPKART** R1 ↑	R2 ↑	RL ↑	BERT-F1 ↑
Extractive	✗	Clustroid	17.92	2.13	10.74	84.27	0.60	0.1	0.60	79.21
	✗	LexRank	22.70	3.10	12.93	83.89	9.66	0.59	6.23	82.62
	✗	QT	21.97	1.66	11.52	83.35	18.83	1.47	10.30	81.65
	✗	SemAE	21.31	1.75	11.30	83.32	-	-	-	-
	✗	$\text{HERCULES}_{\text{EXT}}$	25.49	3.47	12.91	84.01	21.99	1.01	10.16	82.94
Abstractive	✓	CopyCat	16.77	1.57	10.40	83.96	-	-	-	-
	✓	BiMeanVAE	22.12	2.23	12.41	83.85	8.86†	0.70†	6.20†	82.67†
	✓	COOP	24.63	3.04	**14.04**	84.38	9.76†	1.10†	6.71†	82.32†
	✓	$\text{HERCULES}_{\text{ABS}}$	20.21	2.24	11.72	84.37	17.21	0.82	9.76	82.88
INC / HIE	✓	LLAMA-3-8B-8K-INCREMENTAL	25.19	**3.95**	13.35	84.30	38.98	8.56	20.56	86.88
	✓	PHI-3-MINI-3.8B-128K-INCREMENTAL	20.93	2.12	11.18	83.04	35.87	6.58	17.96	85.96
	✓	LLAMA-3-8B-8K-HIERARCHICAL	25.07	3.88	12.73	84.08	33.16	8.30	17.41	86.70
	✓	PHI-3-MINI-3.8B-128K-HIERARCHICAL	24.27	2.81	12.19	84.02	31.09	7.14	14.22	85.56
LLMs	✓	PHI-3-MINI-3.8B-4K	25.34	2.60	12.66	84.16	31.39	5.90	14.62	80.99
	✓	PHI-3-MINI-3.8B-128K	24.14	2.56	12.63	84.36	33.82	7.77	15.62	85.76
	✓	LLAMA-3-8B-8K	26.13	3.12	13.51	84.68	35.35	7.56	17.42	83.77
Ours	✓	XL-OPSUMM(PHI-3-MINI-3.8B-4K)	24.78	2.55	12.72	84.59*	37.71*	6.76	17.83*	86.45*
	✓	XL-OPSUMM(LLAMA-3-8B-8K)	**26.88***	3.52*	13.85*	**85.11***	**39.78***	**8.86**	**21.31***	**87.38***

Table 3: Results on AmaSum, and XL-FLIPKART datasets . *INC/HIE* indicates that the model uses either an Incremental or a Hierarchical approach. Bold and underlined indicate the best and second-best scores. $\star$ indicates p-value < 0.05 on **Wilcoxon Signed-Rank Test** of XL-OPSUMM framework models against their corresponding base LLMs (e.g. XL-OPSUMM(LLAMA-3-8B-8K) vs LLAMA-3-8B-8K). † indicated scores obtained by sampling 8 reviews randomly from test set.

	Model	**GPT-3.5** FL↑	CO↑	**MISTRAL-7B** FL↑	CO↑	BooookScore↑
ABS	$\text{HERCULES}_{\text{ABS}}$	3.76	1.84	4.4	2.36	59.46
INC/HIE	LLAMA-3-8B-8K-INCREMENTAL	4.72	**3.72**	4.56	4.16	70.19
	PHI-3-MINI-3.8B-128K-INCREMENTAL	3.60	2.84	4.16	3.36	57.86
	LLAMA-3-8B-8K-HIERARCHICAL	**4.80**	3.60	**4.92**	**4.44**	71.58
	PHI-3-MINI-3.8B-128K-HIERARCHICAL	4.36	3.48	4.64	3.92	63.12
LLMs	LLAMA-3-8B-8K	4.64	3.44	4.68	4.08	65.06
	PHI-3-MINI-3.8B-4K	4.12	3.40	4.48	3.76	58.97
	PHI-3-MINI-3.8B-128K	4.60	3.56	4.48	4.04	61.86
Ours	XL-OPSUMM(PHI-3-MINI-3.8B-4K)	4.60	3.52	4.44	**4.44**	61.41
	XL-OPSUMM(LLAMA-3-8B-8K)	4.68	3.64	4.56	**4.44**	**85.60**

Table 4: Reference free evaluation on AMASUM Dataset. *INC/HIE* indicates that the model uses either an Incremental or a Hierarchical approach. FL represents the average fluency score across all summaries generated by the model, while CO denotes the average coherency score. Refer to Appendix B for more description of these metrics.

Table 3 presents the results of various models on the AmaSum and XL-FLIPKART datasets. The analysis reveals the effectiveness of the XL-OPSUMM framework, particularly when employed with large language models (LLMs) such as LLAMA-3-8B-8K and PHI3-3-MINI-3.8B-4K.

On the AMASUM dataset, the XL-OPSUMM(LLAMA-3-8B-8K) model outperforms its base and hierarchical variants across all metrics, including R1, R2, RL, and BERT-F1. It achieves the highest scores among all models for R1, RL, and BERT-F1, while being marginally outperformed by its incremental variant in terms of R2. Despite the PHI3-3-MINI-3.8B-4K model exhibiting higher ROUGE scores than the XL-OPSUMM(PHI3-3-MINI-3.8B-4K) model, the Wilcoxon Signed-Rank test indicates that the difference is not statistically significant.

On the XL-FLIPKART testset, the incremental variants of the LLAMA-3 and PHI-3 models outperform their corresponding base and hierarchical counterparts. Notably, when these LLMs are employed within the XL-OPSUMM framework, they surpass the performance of their incremental variants. Specifically, the XL-OPSUMM(LLAMA-3-8B-8K) model achieves the highest or second-highest scores across all metrics, outperforming the previous state-of-the-art models, such as

	Model	GPT-3.5		Mistral-7B		
		FL↑	CO↑	FL↑	CO↑	BooookScore↑
ABS	$\text{HERCULES}_{\text{ABS}}$	4.00	1.64	4.20	2.16	39.56
INC/HIE	Llama-3-8B-8k-Incremental	4.56	3.28	4.52	3.88	70.73
	Phi-3-Mini-3.8B-128k-Incremental	4.04	3.08	4.20	3.76	50.98
	Llama-3-8B-8k-Hierarchical	4.64	3.52	4.44	4.08	64.70
	Phi-3-Mini-3.8B-128k-Hierarchical	4.32	3.36	4.44	4.00	55.59
LLMs	Llama-3-8B-8k	4.60	3.12	4.48	3.72	67.71
	Phi-3-Mini-3.8B-4k	3.76	2.68	4.44	3.44	43.27
	Phi-3-Mini-3.8B-128k	4.47	3.19	4.36	3.44	57.06
Ours	Xl-OpSumm(Phi-3-Mini-3.8B-4k)	**4.68**	**3.68**	**4.72**	**4.16**	66.23
	Xl-OpSumm(Llama-3-8B-8k)	4.48	3.48	4.64	**4.16**	**87.59**

Table 5: Reference-free evaluation on the Xl-Flipkart dataset. *INC/HIE* indicates that the model uses either an Incremental or a Hierarchical approach. FL represents the average fluency score across all summaries generated by the model, while CO denotes the average coherency score. Refer to Appendix B for more description of these metrics.

$\text{HERCULES}_{\text{EXT}}$ and $\text{HERCULES}_{\text{ABS}}$.

The results demonstrate the effectiveness of the Xl-OpSumm framework in leveraging the capabilities of LLMs like Llama-3-8B-8k and Phi3-3-Mini-3.8B-4k for abstractive summarization tasks across diverse datasets like AMASUM and Xl-Flipkart. The framework consistently enhances the performance of these LLMs, enabling them to outperform existing state-of-the-art models.

7.2 Reference Free Evaluation

Traditional reference-based metrics like ROUGE inherently fail to capture the nuances of issues and contradictions within reviews, as demonstrated by prior work ((Bhaskar et al., 2023), (Siledar et al., 2024a)). To address this limitation, we evaluate our framework across two dimensions: `fluency (FL)` and `coherence (CO)`(Appendix B), by prompting GPT-3.5-Turbo and Mistral-7B-32k models using the same method and prompts introduced in Siledar et al. (2024a). We could not evaluate the summaries on `Relevance, Faithfulness, Aspect Coverage, Sentiment Consistency, Specificity` due to their input dependency and the token length limitations of the models under consideration (GPT-3.5-Turbo and Mistral-7B-32k). Additionally, we use `BooookScore` (Chang et al., 2023) to evaluate the coherence of these summaries.

AMASUM Dataset Evaluation

Table 4 presents the reference-free evaluation on the AMASUM dataset. All the LLM-based models outperform the $\text{HERCULES}_{\text{ABS}}$ model across all three metrics. Specifically, Xl-OpSumm(Llama-3-8B-8k) achieves the highest avg[5] `Coherence` score of 4.04 among its Llama-based variants, followed closely by Llama-3-8B-8k-Hierarchical with 4.02. Llama-3-8B-8k-Incremental has an avg score of 3.94. In terms of `Fluency`, Llama-3-8B-8k-Hierarchical leads with an avg score of 4.86, followed by Xl-OpSumm(Llama-3-8B-8k) with an avg score of 4.56. In terms of `BooookScore`, Xl-OpSumm(Llama-3-8B-8k) outperforms all other models with a score of 85.60, followed by Llama-3-8B-8k-Hierarchical which achieved a score of 71.58.

Among the PHI-3 models, Xl-OpSumm(Phi-3-Mini-3.8B-4k) excels with an avg `Coherence` score of 3.98 and an avg `Fluency` score of 4.52. It is closely followed by the Phi-3-Mini-3.8B-128k-Hierarchical model, which has avg scores of 4.5 in `Fluency` and 3.7 in `Coherence`. The Phi-3-Mini-3.8B-128k-Hierarchical model achieved the highest `BooookScore` of 63.12, closely followed by the Phi-3-Mini-3.8B-128k and Xl-OpSumm(Phi-3-Mini-3.8B-4k) models, which scored 61.86 and 61.41 respectively.

Xl-Flipkart Dataset Evaluation

Table 5 displays the reference-free evaluation on the Xl-Flipkart dataset. Models in the Xl-OpSumm framework outperform their Hierarchical and Incremental counterparts. Xl-OpSumm(Llama-3-8B-8k) achieves avg scores of 4.56 in `Coherence` and 3.82 in `Fluency`.

[5]avg: mean of scores given by GPT-3.5 and Mistral-7B models as evaluators

The LLAMA-3-8B-8K-HIERARCHICAL model scores an avg of 3.8 in Coherence and 4.54 in Fluency, while the LLAMA-3-8B-8K-INCREMENTAL model scores an avg of 3.58 in Coherence and 4.54 in Fluency. As observed in the AMASUM dataset, XL-OPSUMM(LLAMA-3-8B-8K) once again outperformed all other models, achieving a BooookScore of 87.59. This time, it was followed by LLAMA-3-8B-8K-INCREMENTAL, which scored 70.73.

A similar trend is observed with the PHI-3-powered models. XL-OPSUMM(PHI-3-MINI-3.8B-4K) achieves the highest avg Coherence score of 3.82 and the highest avg Fluency score of 4.7 and a BooookScore of 66.23 among all PHI-3 models evaluated.

7.3 Qualitative Analysis

Table 1 presents the summaries (Previous Global Summary, Local Summary, and Updated Global Summary) generated for a certain chunk of a Realme 8 product from the XL-FLIPKART dataset using XL-OPSUMM (PHI-3-MINI-3.8B-4K). We observe that aspects such as build quality, Super AMOLED display, and gaming performance are new aspects present in the Local Summary. After referring to the aspect dictionary, aspects like the "display" and "battery life" of the mobile are updated in the global summary from the Previous Global Summary since they have the same sentiment in the aspect dictionary and Local Summary. For aspects like camera quality, there was dissatisfaction in the Previous Global Summary, but satisfaction concerning the back camera and dissatisfaction concerning the front camera in the Local Summary, so they are updated accordingly in the global summary referring to the aspect dictionary as well. Similarly, there was dissatisfaction in the Previous Global Summary for the aspect performance, but it was updated to a positive sentiment by referring to the Local Summary and aspect dictionary.

We also observed that specific information about aspects such as the MTK G95 processor model name, SONY IMX rear camera sensor, and 5000 mAh battery were dropped in some cases. Additionally, we observe a few hallucinations by the model, such as a rating of 4.5 out of 5 stars, which is not present in either the Local Summary or the Previous Global Summary.

7.4 Comparative Analysis

Table 6 shows summaries generated by various models for the Samsung Galaxy F23 5G. We observe that all the LLM-based summaries are coherent. However, the summary generated by $\text{HERCULES}_{\text{ABS}}$ lacks a structured overview and relevance to the product.

While other models successfully extract detailed information about the phone's features, $\text{HERCULES}_{\text{ABS}}$ fails to do so. The Gold (GPT-4) summary stands out with its comprehensive coverage of multiple aspects, including display, battery life, performance, and camera quality, providing a balanced view highlighting both strengths and weaknesses. The LLAMA-3-8B-8K-INCREMENTAL and XL-OPSUMM(LLAMA-3-8B-8K) summaries provide general overviews of user experiences but lack the depth and specific insights found in the Gold summary. The XL-OPSUMM(LLAMA-3-8B-8K) summary, in particular, highlights several positives not mentioned in the LLAMA-3-8B-8K-INCREMENTAL, such as the phone's durability and overall design quality. LLAMA-3-8B-8K-HIERARCHICAL summary, while it is coherent, the length of the summary is very large compared to other model summaries.

8 Summary, Conclusion and Future Work

In this work, we introduce XL-OPSUMM, a scalable framework for opinion summarization that generates summaries incrementally from thousands of reviews. Additionally, we present a new test set, XL-FLIPKART, which contains thousands of reviews per product. Our framework can theoretically scale to process any number of reviews, regardless of the LLM context limit. Experimental results show that our framework outperforms all previous state-of-the-art models and other baselines on two datasets in ROUGE-based evaluations, and achieves higher average scores in reference-free evaluations across three dimensions.

Studies from Siledar et al. (2024b) showed that additional information sources are indeed helpful for opinion summarization task. Inspired from those works, a future direction for our work is to integrate additional sources such as Question-Answers and Product Descriptions into the XL-OPSUMM framework and to analyze their impact in the context of large volumes of reviews.

Limitations

1. Due to budgetary constraints associated with utilizing GPT-4, we have limited the Xl-Flipkart dataset to 25 products, for which we generated summaries using GPT-4. The principal objective of this study is to develop a framework capable of managing extensive contexts efficiently.

2. We could not evaluate the summaries on `Relevance`, `Faithfulness`, `Aspect Coverage`, `Sentiment Consistency`, and `Specificity`. This is because these evaluations depend on the input and the models we used, GPT-3.5-TURBO and MISTRAL-7B-32K, have limitations on token length.

Ethical Considerations

While leveraging GPT-4-TURBO to generate summaries offers significant time and resource savings, we are aware of the potential impact on jobs related to summarizing and analyzing reviews. To address this, we are exploring methods to integrate human oversight with automated processes, striving to balance efficiency with job preservation. Furthermore, users and stakeholders need to understand that these summaries are generated by AI. So we urge the research community to use the XL-FLIPKART test set with caution,

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2020. Unsupervised opinion summarization with content planning. In *AAAI Conference on Artificial Intelligence*.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Somnath Basu Roy Chowdhury, Chao Zhao, and Snigdha Chaturvedi. 2022. Unsupervised extractive opinion summarization using sparse coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1225, Dublin, Ireland. Association for Computational Linguistics.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms.

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Manzil Zaheer, Andrew McCallum, Amr Ahmed, and Snigdha Chaturvedi. 2024. Incremental extractive opinion summarization using cover trees.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-supervised and controlled multi-document opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1646–1662, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Demian Gholipour Ghalandari. 2017. Revisiting the centroid-based method: A strong baseline for multi-document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 85–90, Copenhagen, Denmark. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Aaai*, volume 7, pages 1621–1624.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better topenai human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*. ACM.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Han Jiang, Rui Wang, Zhihua Wei, Yu Li, and Xinpeng Wang. 2023. Large-scale and multi-perspective opinion summarization with diverse review subsets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5641–5656, Singapore. Association for Computational Linguistics.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. Consistsum: Unsupervised opinion summarization with the consistency of aspect, sentiment and semantic. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2023. ChatGPT (August 3 Version). https://chat.openai.com.

Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40:919–938.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, Valencia, Spain. Association for Computational Linguistics.

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023. Synthesize, if you do not have: Effective synthetic dataset creation strategies for self-supervised opinion summarization in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13480–13491, Singapore. Association for Computational Linguistics.

Tejpalsingh Siledar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, and Nikesh Garera. 2024a. One prompt to rule them all: Llms for opinion summary evaluation.

Tejpalsingh Siledar, Rupasai Rangaraju, Sankara Sri Raghava Ravindra Muddu, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, Nikesh Garera, Swaprava Nath, and Pushpak Bhattacharyya. 2024b. Product description and qa assisted self-supervised opinion summarization.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction tuning for few-shot aspect-based sentiment analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Ke Wang and Xiaojun Wan. 2021. TransSum: Translating aspect and sentiment embeddings for self-supervised opinion summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 729–742, Online. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 47–57, San Diego, California. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

A Global Summary Updation

When we process a chunk of reviews, they may have certain aspects that are not present in the previous chunks or may have information about the same aspects that conflict with the opinions from the previous chunks. Typically GLOBAL SUMMARY represents important information from the previous chunks and the LOCAL SUMMARY represents important information from the current chunk. When updating information in GLOBAL SUMMARY using LOCAL SUMMARY, we handle the below 2 cases i.e. having new aspects in the LOCAL SUMMARY and conflicting opinions between LOCAL SUMMARY and GLOBAL SUMMARY with the help of ASPECT DICTIONARY

a. New Aspects in the LOCAL SUMMARY: In the case of a new aspect, we check the *Aspect Dictionary* for the majority sentiment of that aspect. We only update GLOBAL SUMMARY with new aspect information if the sentiment of that aspect in the LOCAL SUMMARY and the ASPECT DICTIONARY matches. By doing this, we are making sure that the summary stays faithful to that aspect.

b. Conflicting Opinions about an aspect between GLOBAL and LOCAL SUMMARIES: In such cases, we again refer to the aspect's majority sentiment from the ASPECT DICTIONARY. If it matches with sentiment in LOCAL SUMMARY, we update the GLOBAL SUMMARY with the corresponding information from the LOCAL SUMMARY, else we leave the GLOBAL SUMMARY as it is.

We embed all this information in a detailed prompt with a one-shot example and feed it to the LLM to update the GLOBAL SUMMARY.

Gold (GPT-4)
The Samsung Galaxy F23 5G offers a decent display, good battery life, and solid performance with the Snapdragon 750G at a competitive price. While the display is a TFT panel, it performs adequately for most users. The device's rear camera delivers satisfactory daylight images, but the front camera and low-light performance are average. Users appreciate the 120Hz refresh rate, although it experiences some inconsistency. Despite featuring a robust build and offering future-proof 5G compatibility, some customers have issues with fingerprint scanner reliability, adaptive brightness, and occasional software glitches. Notably, the phone comes without a charger, adding an extra expense for buyers. The device handles moderate gaming and day-to-day tasks well, with minor heat issues during extended use. While it includes Android 12 and promises good security features, users are concerned about the frequency and reliability of updates from Samsung. Overall, the phone is seen as a good, budget-friendly option, particularly for those prioritizing non-Chinese brands, despite some shortcomings in camera capabilities and accessory inclusion.
HERCULES$_{\text{abs}}$
Will not power on. This camera is okay. Good performance. No more Cable bills! Idea. Tried multiple times. Did not need any Tech support. Doesn't hold a charge. Works fine. This thing is a beast. Very happy with this purchase. Good luck.
LLAMA-3-8B-8K-INCREMENTAL
Here is the updated summary: The Samsung Galaxy F23 has received mixed reviews. While some praise its performance, camera, and display, others have reported issues with battery life, camera quality, and heating. The phone lacks a charger and charging cable in the box, a major drawback. Some users have reported issues with low-light camera performance, fast-draining battery, and auto brightness. However, the phone's display is good, and it has a smooth gaming experience. Overall, it's a good budget option, but may not meet premium device expectations.
LLAMA-3-8B-8K-HIERARCHICAL
Here is the merged summary: **Opinion Summary:** The Samsung Galaxy F23 5G is a decent budget smartphone that offers a mix of pros and cons. While many users praise its performance, display, and camera quality, some have reported issues with battery life, with some experiencing rapid draining even with moderate usage. The phone's charger is not included in the box, which may be an inconvenience for some users. The camera quality is a mixed bag, with some users praising it and others reporting poor performance in low-light conditions. The phone's 5G connectivity is a major selling point, although some users are disappointed that it does not support JIO 5G yet. Additionally, ...
XL-OPSUMM (LLAMA-3-8B-8K)
The Samsung phone has received mixed reviews. Users praise its camera quality, performance, and durability. However, some users have reported issues with battery drainage, overheating, and poor charging speed. Some users also experienced connectivity issues, sound quality issues, and poor network quality. The phone's display and design are also a subject of debate, with some finding it to be good, while others think it's average. Overall, opinions on the phone's value for money and performance vary, with some finding it a good budget option and others considering it a waste of money.

Table 6: Comparative Analysis. Summaries generated by various models on a product from the XL-FLIPKART dataset.

B Various Metrics Used in this Work

ROUGE-1 (R1)(Lin, 2004) measures the overlap of unigrams (single words) between the generated summary and the reference summary. It gives an indication of how many individual words from the reference summary are captured in the generated summary.

ROUGE-2 (R2)(Lin, 2004) measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary. It provides insight into how well the generated summary preserves the sequence of word pairs from the reference summary.

ROUGE-L (RL)(Lin, 2004) calculates the longest common subsequence (LCS) between the generated summary and the reference summary. It captures the longest sequence of words that appear in both summaries in the same order, providing a measure of the overall structural similarity between the summaries.

BERT-F1(Zhang et al., 2019) uses BERT, a pre-trained language model, to evaluate the similarity between the generated summary and the reference summary. BERTScore calculates precision, recall, and F1 score by comparing the contextual embeddings of words in both summaries, providing a more nuanced measure of semantic similarity than simple n-gram overlap.

FLUENCY (FL)(Siledar et al., 2024a) assesses the quality of a summary in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure. A fluent summary should be free of errors, and easy to read, follow, and comprehend. Annotators were given specific guidelines on how to penalize summaries based on their fluency levels.

COHERENCE (CO)(Siledar et al., 2024a) evaluates the overall quality of the sentences in a summary. A coherent summary should be well-structured and well-organized, forming a logical and connected body of information rather than just a collection of related sentences.

BOOOOKSCORE(Chang et al., 2023) evaluates the coherence of summaries by prompting large language models (LLMs) to identify eight types of errors in each sentence. These errors include entity omission, event omission, causal omission, discontinuity, salience, language issues, inconsistency, and duplication. This metric is both reference-free and source-free.

C Other Baselines

This section contains baselines that are not discussed in section 6.1

Oracle represents the extractive upper bound computed by selecting input sentences with the highest R1 compared to the gold summary.

Random represents selecting random reviews from the input as a lower bound.

LexRank (Erkan and Radev, 2004) represents selecting the most salient sentences from the input by using BERT encodings to encode the sentences.

QT (Angelidis et al., 2021) represents using vector quantization to map sentences to a discrete encoding space, then generates extractive summaries by selecting representative sentences from clusters.

SemAE (Basu Roy Chowdhury et al., 2022) extends QT, relaxing the discretization and encoding sentences as mixtures of learned embeddings.

CopyCat (Bražinskas et al., 2020) uses a hierarchical variational autoencoder that learns a latent code of the summary.

D Examples

We present the example summaries of a few baselines, XL-OPSUMM(LLAMA-3-8B-8K) and XL-OPSUMM(PHI-3-MINI-3.8B-4K) in the Table 7.

Gold Summary

After evaluating numerous reviews, the Motorola G60 appears to be a robust mid-range smartphone offering consistent performance. Buyers have appreciated its large LCD display, smooth software experience with clean stock Android, and reliable battery life that comfortably lasts for over two days. The Snapdragon 732G chipset ensures lag-free gaming and smooth daily operation. While the device has a considerable weight and lacks an AMOLED screen, most users find these drawbacks negligible compared to the overall performance and build quality. The camera setup, though slightly slow in processing, provides satisfactory image quality in various settings. Critical points include the absence of stereo sound and fast charging, with the device taking around 2.5 hours for a full charge. Overall, the Moto G60 is considered a good value for its price, particularly praised for its user interface and build, making it suitable for users prioritizing functionality over high-end specifications.

HERCULES$_{\text{ABS}}$

Very heavy. This camera does not disappoint. The battery lasts a long time. Doesn't feel cheap at all. Jim Does not come with a case. Worst product ever. Great sound and battery life. The battery last a long time. The sound quality is not that great. Very helpful. This camera is much better than the older model. Bought this for my wife for Christmas. Low light performance is great.

LLAMA-3-8B-8K

Here is a comprehensive opinion summary of the Motorola Moto G60:**Performance:** The phone has a good performance, with most users reporting smooth operation and no lag. **Display:** The display is a highlight of the phone, with many users praising its brightness, clarity, and refresh rate. **Camera:** The camera is a mixed bag, with some users reporting good quality images and others finding it to be below average. The front camera is generally praised, while the rear camera is criticized for its poor performance. **Battery:** The battery life is excellent, with many users reporting that it lasts for two days with moderate use. **Design:** The phone's design is a subject of debate, with some users finding it too bulky and others liking its sleek design. **Software:** The phone runs on stock Android, which is a major plus for many users. However, some users report occasional software issues and lag. **Value for money:** The phone is generally considered to be a good"

PHI-3-MINI-3.8B-4K

Based on these reviews, it seems like the sentiment towards the Motorola Moto G60 is mixed. Reviewers appreciate its display quality, camera performance, and value for money. However, they also express dissatisfaction with issues related to heating, camera functionality, and app performance, wh'

LLAMA-3-8B-8K-INCREMENTAL

Here is the updated summary: The Motorola G60 has received mixed reviews, with some users praising its performance, battery life, and display, while others have been disappointed with the camera quality. The phone's camera has been criticized for its poor low-light performance and some users have reported poor results with the selfie camera. The phone's display, battery life, and performance have been praised, but the camera quality is a major letdown. Despite this, the phone's performance is generally good, with no major lag or issues reported. The phone's battery life is also a highlight, with some users reporting up to 2.5 days of moderate use. However, the phone's charging time is slow and some users have reported issues with the camera app. Overall, the Motorola G60 is a good option for those looking for a budget-friendly device with good performance and battery life, but may not be the best choice for those prioritizing camera quality.

PHI-3-MINI-3.8B-128K-INCREMENTAL

Updated Summary: Users have noted the Moto G60's sturdy construction, impressive battery life, and vibrant display, along with a solid stock Android experience. However, several users have criticized the phone for its slow camera, particularly in low light, heating issues during operation and charging, and s, and with with and and and and with. There. This such. And. and.s. and and and and and some for and with withs. It and ands. and and the, and with and with and and like and and and and ands and and and and and and such such to such like like. buts. This. like.s.s like such ands like such. like such. However. and and, and . with with with, like. and and like. and and such like and such like like such. with with such, and like. and. However. However, but. Thiss. However withs.s, .s and to confirm, and. I. However.s.. And. There. and and. in ands ands. and. and. with with, and, and and with, a

LLAMA-3-8B-8K-HIERARCHICAL

Here is the merged summary: The Motorola Moto G60 has received mixed reviews from customers, with both praise and criticism for its various features. On the positive side, many users appreciate its excellent battery life, good display, decent camera quality, and smooth performance. The phone's stock Android experience, fast charging, and good value for money are also notable strengths. However, some users have been disappointed with the camera quality, particularly in low-light conditions, and slow charging. Additionally, some have reported issues with hanging, overheating, poor camera performance, touch response, ghost touches, heating, and Wi-Fi connectivity problems. Some common praises include: Good performance and ...

PHI-3-MINI-3.8B-128K-HIERARCHICAL

The Motorola G60 phone has garnered a mix of reviews, with users praising its commendable performance, long battery life, clean Android experience, compact design, and decent camera capabilities. Notably, the phone is appreciated for offering value for money within its price range. However, the reviews also consistently mention several areas of dissatisfaction. These include the display size and quality, particularly with the camera, in addition to software-related issues like heating during charging and app functionality. Speaker quality and the challenge in finding appropriate...

XL-OPSUMM(PHI-3-MINI-3.8B-4K)

The sentiment expressed in these reviews on the phone is mixed, with users appreciating the phone's performance, display, battery life, and clean UI. However, there are concerns about the heavy weight, poor camera quality (especially in natural light), and issues with the charging speed. Users also find the touch response to be slower than expected and note concerns about the phone's overheating during charging. Some users are satisfied with the phone's features and performance, while others find it not suitable for photography enthusiasts. The phone seems to offer good value for money, but it falls short in terms of camera and display quality compared to competitors. It is recommended for stock Android lovers and those prioritizing performance over photography. Overall, it's perceived as a great purchase in a big sale but advises caution due to some drawbacks."

XL-OPSUMM(LLAMA-3-8B-8K)

The Motorola Moto G60 has received mixed reviews, with some users praising its battery life, camera, and stock Android experience. However, many others have reported issues with the cameraś performance, particularly in low-light conditions. Some users have also experienced heating problems, slow Wi-Fi connectivity, and poor customer service. Additionally, users have praised the phoneś display quality and value for money. Overall, the phone seems to be a decent option for those looking for a budget-friendly device, but itś essential to set realistic expectations and weigh the pros and cons before making a purchase.

Table 7: Summaries generated by various models about Motorola G60 smart phone from XL-FLIPKART dataset. Gold Summary is summary generated by GPT-4-Turbo model.