Product Description and QA Assisted Self-Supervised Opinion Summarization

Tejpalsingh Siledar*♣, Rupasai Rangaraju*♣, Sri Raghava Muddu*♣,
Swaprava Nath♣, Pushpak Bhattacharyya♣,
Suman Banerjee♠, Amey Patil♠, Sudhanshu Shekhar Singh♠,
Muthusamy Chelliah♠, Nikesh Garera♠
♣Computer Science and Engineering, IIT Bombay, India,
♠Flipkart, India
{tejpalsingh, rupasai, sriraghava, swaprava, pb}@cse.iitb.ac.in

Abstract

In e-commerce, opinion summarization is the process of summarizing the consensus opinions found in product reviews. However, the potential of additional sources such as *product description* and *question-answers (QA)* has been considered less often. Moreover, the absence of any supervised training data makes this task challenging. To address this, we propose a novel synthetic dataset creation (SDC) strategy that leverages information from reviews as well as additional sources for selecting one of the reviews as a pseudo-summary to enable supervised training. Our Multi-Encoder Decoder framework for Opinion Summarization (**MEDOS**) employs a separate encoder for each source, enabling effective selection of information while generating the summary. For evaluation, due to the unavailability of test sets with additional sources, we extend the Amazon, Oposum+, and Flipkart test sets and leverage ChatGPT[1] to annotate summaries. Experiments across nine test sets demonstrate that the combination of our SDC approach and MEDOS model achieves on average a $\mathbf{14.5}\%$ improvement in ROUGE-1 F1 over the SOTA. Moreover, comparative analysis underlines the significance of incorporating additional sources for generating more informative summaries. Human evaluations further indicate that MEDOS scores relatively higher in *coherence* and *fluency* with 0.41 and 0.5 (-1 to 1) respectively, compared to existing models. *To the best of our knowledge*, we are the first to generate opinion summaries leveraging additional sources in a self-supervised setting.

1 Introduction

In the e-commerce domain, reviews play a vital role in making informed decisions. However, due to the recent proliferation of online reviews, going through all the product reviews before making a decision is challenging. Opinion summarization provides a solution by summarizing the opinions presented in the reviews (Hu and Liu, 2006; Wang and Ling, 2016; Angelidis and Lapata, 2018). However, text summarization (Nallapati et al., 2016; See et al., 2017; Liu and Lapata, 2019) usually contains reference summaries which are very difficult to obtain at a large scale for opinion summarization. As a result, recent studies (Bražinskas et al., 2020; Elsahar et al., 2021) enable self-supervision by curating synthetic pairs out of review corpus by sampling one of the reviews as a pseudo summary and considering the remaining reviews as the input.

MultimodalSum
I bought this product to scan my negatives. It does not work with Windows XP. I have tried to contact the company several times and have not received a response. I am very disappointed in the product. I would not recommend it to anyone.
Our Model (MEDOS)
I purchased the **VuPoint FS-C1-VP Film and Slide Digital Converter** to scan my **35mm** film and slide negatives. It is not compatible with Windows XP. The software does not work with Windows 7 or 8. I have tried to contact the company and they do not respond to my emails. I would not recommend this product to anyone.

Table 1: MultimodalSum vs. MEDOS generated summary for a product from the Amazon test set. Information assisted from product description and question-answers are in **bold** and underline respectively. Our model is able to capture essential information from the product description and question-answers, not found in reviews. This makes our model-generated summaries more informative while still retaining the consensus opinions from reviews as evident in the above example.

Motivation In e-commerce, users' opinions are expressed through various sources such as product ratings, reviews, review upvotes and downvotes, and question-answers. Additionally, for each product, description, product specification, product im-

*Equal contribution.
[1]https://chat.openai.com/ (gpt-3.5 August 3 version)

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2315–2332

ages, price, etc. are present as well. Considering such additional sources apart from reviews is vital in generating opinion summaries that are well-rounded and informative. Specifically, descriptions offer nuanced details about various aspects, while question-answers provide additional perspectives on specific queries, both of which can be valuable. Table 1 shows an example of the influence of product description and question-answers. However, acquiring annotated training datasets proves expensive and impractical as the number of sources increases. This makes it essential to devise effective synthetic dataset creation strategies that enable supervised training of models using multiple sources.

Problem Statement We propose a novel synthetic dataset creation approach that uses additional sources such as *product description* and *question-answers (QA)* along with reviews for generating synthetic quadruplets of the form *{input reviews, description, question-answers, pseudo-summary}* to enable end-to-end supervised training. A multi-encoder decoder model for opinion summarization (**MEDOS**) to effectively select information from either product description or question-answers while summarizing reviews. For evaluation, due to the unavailability of test sets that have annotated summaries written considering such additional sources (except for Flipkart (Siledar et al., 2023b)), we extend the available e-commerce test sets by including these additional sources and leveraging ChatGPT (OpenAI, 2023) to annotate (Gilardi et al., 2023; Huang et al., 2023) summaries.
Input: *Reviews, Description, Question-Answers*
Output: *Opinion Summary*

Our contributions are:

1. A novel synthetic dataset creation (SDC) approach that enables supervised training in the presence of additional sources without the need for any annotated training datasets. We propose a Multi-Encoder Decoder framework for Opinion Summarization (**MEDOS**)[2] to effectively fuse information from *reviews, product description,* and *question-answers (QA)* (Section 3, 4 & 5). *To the best of our knowledge*, we are the first to do multi-source self-supervised opinion summarization.

2. Extensions to e-commerce test sets namely Amazon (Bražinskas et al., 2020) and Oposum+ (Amplayo et al., 2021) to include additional sources. For comparison, we extend: **Amazon**, **Oposum+**, and **Flipkart** by curating six new test sets: Amazon R, Amazon RDQ, Oposum+ R, Oposum+ RDQ, Flipkart R, and Flipkart RDQ leveraging ChatGPT to annotate summaries. We extend the test sets to contain **662 opinion summaries** across six curated test sets (Section 6.2, Table 2).

3. Experimental demonstrations of our SDC approach and MEDOS model in outperforming the SOTA model on **nine test sets** on average by $\mathbf{14.5}\%$ in ROUGE-1 F1 (Section 7).

4. Comparative and qualitative analysis indicating the importance of sources such as *product description* and *question-answers* in generating more informative summaries compared to existing models (Section 7, Table 4 & 5).

2 Related Work

Self-supervised Opinion Summarization. Recent approaches use self-supervision by considering one of the reviews as a pseudo-summary. Bražinskas et al. (2020) randomly selected N reviews per entity to construct N pseudo-summary, reviews 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 sample pseudo-summaries. Elsahar et al. (2021) selected reviews similar to a randomly sampled pseudo-summary as input reviews, based on TF-IDF cosine similarity. Wang and Wan (2021) aimed at reducing opinion redundancy and constructed highly relevant reviews pseudo-summary pairs by learning aspect and sentiment embeddings to generate relevant pairs. Im et al. (2021) used synthetic dataset creation strategy similar to Bražinskas et al. (2020) and extended it to multimodal version. Ke et al. (2022) captured the consistency of aspects and sentiment between reviews and pseudo-summary using constrained sampling. Siledar et al. (2023a) use lexical and semantic similarities for creating synthetic datasets. Our work is most similar to Elsahar et al. (2021) in using cosine similarity to select input reviews and pseudo-summary pairs. However, we use review embeddings to compute similarity instead of TF-IDF scores. Additionally, our pseudo-summary

[2]Code and data: https://github.com/tjsiledar/MEDOS

	Original			Extended (Ours)					
	Amazon	Oposum+	Flipkart	Amazon GPT-R	Oposum+ GPT-R	Flipkart GPT-R	Amazon GPT-RDQ	Oposum+ GPT-RDQ	Flipkart GPT-RDQ
#domains	4	6	3	4	6	3	4	6	3
#test set	32	30	145	32	30	145	32	30	145
#reviews/product	8	10	10	8	10	10	8	10	10
#summaries/product	3	3	1	**3**	**3**	**1**	**3**	**3**	**1**
#summaries	96	90	145	**96**	**90**	**145**	**96**	**90**	**145**
#descriptions	-	-	-	-	-	-	21	17	145
#question-answers	-	-	-	-	-	-	11	10	145

Table 2: Statistics for original and extended test sets. GPT-R indicates the use of *reviews* whereas GPT-RDQ indicates the use of *reviews*, *description*, and *question-answers* to generate summaries using ChatGPT. **Bold** represents our contributions. In the respective extended versions, reviews are the same as the original.

selection considers additional sources such as *product description* and *question-answers* as well. Our synthetic dataset creation strategy ensures that the pseudo-summary selection is highly relevant to all our input sources. Recent opinion summarization systems (Bhaskar et al., 2023; Hosking et al., 2023) include a large number of reviews. However, we limit our work to a fixed number of reviews to enable a fair comparison with previous approaches.

Additional sources for Opinion Summarization. Zhao and Chaturvedi (2020) used aspects identified from product description to perform extractive aspect-based opinion summarization. Li et al. (2020) proposed a supervised multimodal summarization model to effectively generate summaries using reviews, product image, product title, and product details. Im et al. (2021) proposed a self-supervised multimodal training pipeline to generate summaries using reviews, images, and metadata. Siledar et al. (2023b) did supervised opinion summarization using simple rules to generate summaries separately in the form of verdict, pros, cons, and additional information using reviews, description, specifications, and question-answers. Our work takes inspiration from Im et al. (2021) to utilize a multi-encoder framework to effectively fuse information from various sources. However, where additional sources are all text, our approach of forming highly relevant synthetic pairs using additional sources helps in capturing relevant information. Also, our approach differs from Siledar et al. (2023b) in training models in an end-to-end fashion without the aid of supervised summaries.

3 Problem Formulation

Preliminaries. For a specific product or an entity, $R = \{r_1, ..., r_N\}$ is the set of N reviews, D represents the product description, and $Q = \{q_1, ..., q_M\}$ represents a set of M question-answer pairs such that q_i represents the i^{th} concatenated question and its corresponding answer.

Opinion Summarization. The task of opinion summarization is to generate an opinion summary s given a set of reviews R for an entity (eg. product or business). Rush et al. (2015) defined the task of abstractive summarization as:

$$s^* = \underset{s}{\text{argmax}}\ g(s, R), \tag{1}$$

$$g(s, R) = \log\ p(s|R; \theta), \tag{2}$$

$$\approx \sum_{i=0}^{J-1} \log\ p(s_{i+1}|s_w, R; \theta), \tag{3}$$

where g is a scoring function defined as a conditional log probability of the summary given the input, $s_w = s_{[i-w+1,...,i]}$ for a window size w, θ is the neural network parameters, and $|s| = J$. For opinion summarization, the input is a review set R and the output is the opinion summary s. The conditional probability can be modeled using Transformers (Vaswani et al., 2017) as:

$$p(s_{i+1}|s_w, R; \theta) \propto \rho(\text{FFN}(\text{C-Attn}(\mathbf{a_R}, \mathbf{e_{s_w}}))), \tag{4}$$

$$\mathbf{a_R} = \text{S-Attn}(\text{Enc}(R)),\ \mathbf{e_{s_w}} = \text{Emb}(s_w), \tag{5}$$

where ρ is the softmax function, FFN is the feed-forward network, C-Attn is the cross-attention network, S-Attn is the self-attention network, Enc is the encoder, and Emb is the embedding layer.

Additional Sources. Under the presence of additional sources such as product description and question-answers, the equations for modeling ab-

Algorithm 1 SDC using Additional Sources

Require: Reviews R, $\mathbf{e_R} \in \mathbb{R}^{N \times d}$, product description D, $\mathbf{e_D} \in \mathbb{R}^{1 \times d}$, and question-answer pairs Q, $q \in Q$, $\mathbf{e_q} \in \mathbb{R}^{1 \times d}$ for a product. Functions sim, $diag$, and $mean$.

Initialize $Z = []$
for each product **do**
  $M \leftarrow diag(sim(\mathbf{e_R}, \mathbf{e_R}), 0)$ $\{\in \mathbb{R}^{N \times N}\}$
  $ds \leftarrow sim(\mathbf{e_R}, \mathbf{e_D})$ $\{\in \mathbb{R}^{N \times 1}\}$
  for $q \in Q$ **do**
    $qs \mathrel{+}= sim(\mathbf{e_R}, \mathbf{e_q})$ $\{\in \mathbb{R}^{N \times 1}\}$
  end for
  $qs \leftarrow mean(qs)$ $\{\in \mathbb{R}^{N \times 1}\}$
  $ss \leftarrow \lambda_1 \cdot ds + \lambda_2 \cdot qs$
  $R_p \leftarrow$ top-p reviews using ss
  for $r \in R_p$ **do**
    $T \leftarrow$ top-k reviews for r using M
    $Z.insert(\{T, D, Q, r\})$
  end for
end for
Return Z

stractive summarization can be written as:

$$s^* = \underset{s}{\text{argmax}}\ g(s, R, D, Q), \quad (6)$$

$$g(s, R, D, Q) = \log\ p(s|R, D, Q; \theta), \quad (7)$$

$$\approx \sum_{i=0}^{J-1} \log\ p(s_{i+1}|s_w, R, D, Q; \theta), \quad (8)$$

Using transformers, this can be modeled as:

$$p(s_{i+1}|s_w, R, D, Q; \theta) \propto \rho(\text{FFN}(\text{C-Attn}(\mathbf{a_f}, \mathbf{e_{s_w}}))), \quad (9)$$

$$\mathbf{e_{s_w}} = \text{Emb}(s_w), \quad (10)$$

where $\mathbf{a_f}$ is the fused attention. We propose a Multi-Encoder Decoder Framework- MEDOS (Section 5, Figure 1) to create fused attention $\mathbf{a_f}$ (Eq. 11).

4 Synthetic Dataset Creation (SDC)

Before discussing the details of our framework, we formalize the synthetic dataset creation process used to train these models. In the absence of supervised datasets, most recent approaches (Bražinskas et al., 2020; Im et al., 2021) resort to self-supervision wherein {input reviews, pseudo-summary} pairs are constructed.

Following Bražinskas et al. (2020), we can assume that a review $r \in R$ can serve as a summary for a set of reviews $T \subseteq R - \{r\}$. This lets us create training points (T, r) i.e. {input reviews, pseudo-summary}, similar to what the model will experience during inference. T is fixed to size k, enabling comparison with existing works.

However, in the presence of additional sources such as product description D and question-answer pairs Q, we slightly modify this definition. Instead of synthetic pairs, we construct synthetic quadruplets of the form: {input reviews, product description, question-answers, pseudo-summary}.

Algorithm 1 details the process of generating synthetic quadruplets. We generate multiple such quadruplets out of reviews R, product description D, and question-answer pairs Q for a specific product. The overall idea for synthetic dataset creation is to choose relevant quadruplets for training. Here we define relevance as the quadruplet that best aids our model in learning the task of opinion summarization using multiple sources.

The intuition is to first select a pseudo-summary r that is the closest to both D and Q. We measure closeness in terms of cosine similarity sim between their embeddings (SBERT (Reimers and Gurevych, 2019)). This selection ensures that the pseudo-summary r contains information relevant to both D and Q so that the model learns to pick information from these two sources as well during training. Next, using the pseudo-summary r selected, we look for its closest k set of reviews that can act as its input reviews set T, which ensures that the model learns the task of summarization.

More formally, we first compute a matrix $M \in \mathbb{R}^{N \times N}$ by computing cosine similarity between embeddings of each review pair (r_a, r_b) where $r_a, r_b \in R$. We make all the diagonals of M as zero to remove self-comparisons using $diag$ function. Next, we compute $ds \in \mathbb{R}^{N \times 1}$ by computing cosine similarity between the embeddings of each review r_a and D. We also compute $qs \in \mathbb{R}^{N \times 1}$ by computing cosine similarity between the embeddings of each review r_a and all $q \in Q$ and taking a $mean$ of it respectively. Finally, we compute $ss \in \mathbb{R}^{N \times 1}$ as $\lambda_1 \cdot ds + \lambda_2 \cdot qs$ where λ_1, λ_2 are parameters set to 0.5 for our experiments. We select $R_p \subseteq R$ reviews for forming p synthetic quadruplets by taking the top-p scores from ss. For each review $r \in R_p$, we get the top-k reviews T from $R - \{r\}$ using scores corresponding to the review r from M. This lets us form synthetic quadruplet

					Amazon			Amazon GPT-R			Amazon GPT-RDQ		
abs?	**Model**	**R**	**D**	**Q**	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑
✗	Random	✓	✗	✗	27.86	3.87	16.68	20.69	1.56	12.55	18.83	1.45	12.03
✗	Oracle	✓	✗	✗	44.47	13.83	30.85	33.69	6.04	22.88	31.83	5.77	22.04
✗	Clustroid	✓	✓	✓	29.27	4.41	17.78	22.74	2.16	14.03	21.31	2.57	13.38
✗	LexRank	✓	✓	✓	29.46	5.53	17.74	22.82	3.08	13.77	19.30	4.31	12.90
✗	QT	✓	✓	✓	34.04	7.03	18.08	23.01	2.48	12.05	21.78	3.25	12.36
✓	CopyCat	✓	✗	✗	31.97	5.81	20.16	20.09	1.79	12.94	20.54	1.94	13.85
✓	PlanSum	✓	✗	✗	32.87	6.12	19.05	20.49	1.76	12.44	19.09	1.58	12.02
✓	ConsistSum	✓	✗	✗	33.32	5.94	**21.41**	-	-	-	-	-	-
✓	MultimodalSum	✓	✓	✗	34.19	7.05	20.81	21.43	1.58	13.20	20.39	2.08	12.83
✓	TransSum	✓	✗	✗	34.23	7.24	20.49	-	-	-	-	-	-
✓	COOP	✓	✗	✗	**36.57**	7.23	21.24	-	-	-	-	-	-
✓	T5-concat	✓	✓	✓	28.04	4.46	16.39	21.28	**2.57**	13.00	20.61	2.72	13.33
✓	BART-concat	✓	✓	✓	32.35	6.49	19.78	22.32	2.27	13.74	21.75	2.39	13.57
✓	**MEDOS**	✓	✓	✓	34.63	**7.48**	20.97	**23.92***	2.27*	**14.69***	**25.44***	**4.16***	**16.45***

Table 3: Results on Amazon test set and its extensions. R, **D**, **Q** indicate the presence of *reviews*, *description*, and *question-answers* respectively in the input. *abs?* indicate abstractive systems. **Bold** and underline indicate best and second-best scores using abstractive systems. * indicates pvalue < 0.05 on paired t-test against MultimodalSum. Overall our combination of SDC approach and MEDOS outperforms baselines across all three test sets.

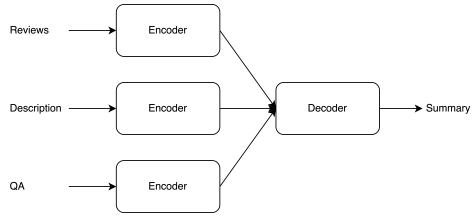

Figure 1: Framework of our MEDOS model that takes *reviews*, *description*, and *question-answers (QA)* as the input. During inference, the model generates a summary whereas during training the model uses pseudo-summary obtained through SDC process for learning.

instances such as $\{T, D, Q, r\}$ for model training.

5 Model Framework (MEDOS)

Figure 1 represents our multi-encoder framework, where each source passes through its separate encoder to generate separate attentions: $\mathbf{a_R}$ = S-Attn(Enc(T)), $\mathbf{a_D}$ = S-Attn(Enc(D)), and $\mathbf{a_Q}$ = S-Attn(Enc(Q)). The fused attention $\mathbf{a_f}$ is then computed as:

$$\mathbf{a_f} = \mathbf{a_R} + \alpha \odot \mathbf{a_D} + \beta \odot \mathbf{a_Q} \quad (11)$$

where $\odot$ represents element-wise multiplication, α and β act as gates regulating the flow of information from product description and question-answers, computed as: $\alpha = \phi([\mathbf{a_R}; \mathbf{a_D}]\mathbf{W}_\alpha)$ and $\beta = \phi([\mathbf{a_R}; \mathbf{a_Q}]\mathbf{W}_\beta)$ where $\mathbf{W}_\alpha, \mathbf{W}_\beta$ are learned parameters and $\phi(\mathbf{x}) = \text{RELU}(\tanh(\mathbf{x}))$ is the activation, following Im et al. (2021).

6 Experiments

6.1 Datasets

We conducted experiments on: Amazon (He and McAuley, 2016; Bražinskas et al., 2020), Oposum+ (Amplayo et al., 2021), and Flipkart (Siledar et al., 2023b). Statistics are in Table 2. Using our SDC strategy, we created 387k and 313k instances from the Amazon and Oposum+ respectively to enable supervised training. Due to the unavailability of review data in the case of Flipkart, we used the Amazon data to train models. Refer **Appendix F**.

6.2 Test Dataset Extension

In the absence of any test sets that contain additional sources, we extended Amazon, Oposum+, and Flipkart to contain such sources and leveraged ChatGPT to annotate summaries using reviews and additional sources as input, amounting to 662 opinion summaries in total. Statistics for the extended versions of the test sets are in Table 2. For extensions, we obtain the additional sources (except for Flipkart) from the Amazon data (He and McAuley, 2016). We leverage ChatGPT as our annotator following recent works (Gilardi et al., 2023; Huang et al., 2023). For each test set, we curated: **GPT-R**, in which summaries are generated using only reviews, and **GPT-RDQ**, in which summaries are generated using reviews, description, and question-answers. We investigated multiple prompts before finalizing the best one (**Appendix C**). We em-

Product 1	Product 2	Product 3	Product 4
I purchased the **VuPoint FS-C1-VP Film and Slide Digital Converter** to scan my **35mm film** and slide negatives. It is not compatible with Windows XP. The software does not work with Windows 7 or 8. I have tried to contact the company and they do not respond to my emails. I would not recommend this product to anyone.	The **Marpac TSC 330 Travel Sound Conditioner** is a great little machine. It is small enough to travel with, but big enough to be used at home. The sound quality is great and it is easy to use. The only thing I don't like about it is that it doesn't have a volume control.	The **Sony Speaker Dock** is a great product. The sound is great and the remote control works great. The only thing I don't like about it is that it doesn't charge my iphone 4s. I have to buy an adaptor for that.	The **Opteka HG-1 Heavy-Duty Aluminum Ultra Hand-Grip Handheld Stabilization System for DSLR and Video Cameras** is a great product. I use it with my Nikon Coolpix L820 and it works great. It is a little heavy, but that is to be expected for a small camera.

Table 4: Qualitative Analysis. MEDOS generated summaries for four different products from the Amazon test set utilizing *reviews, description, and question-answers*. Information assisted by the product description is indicated in **bold**, whereas those assisted from the question-answers are underlined.

ployed three professionals to evaluate the annotation quality on *informativeness, faithfulness, coherence, conciseness,* and *fluency* using a 5-point scale. Statistics are in Table 15. The Inter-Rater Reliability computed using Fleiss' Kappa was $0.23, 0.41$ and 0.42 for human-annotated, GPT-R, and GPT-RDQ summaries which are considered *fair, moderate,* and *moderate* agreement respectively (Landis and Koch, 1977). Refer to **Appendix D** & **I**.

6.3 Baseline Models

Extractive Approaches. *Random* selects a random review from the input as a lower bound. *Oracle* is the extractive upper bound computed by selecting input sentences with the highest R1 to gold summary. *Clustroid* (Bražinskas et al., 2020) selects the review with the highest RL score with respect to other reviews. *LexRank* (Erkan and Radev, 2004) selects the most salient sentences from the input using BERT (Devlin et al., 2019) encodings to represent sentences. *QT* (Angelidis et al., 2021) represents opinions in quantized space.

Abstractive Approaches. *CopyCat* (Bražinskas et al., 2020) is a hierarchical variational autoencoder that learns a latent code of the summary. *PlanSum* (Amplayo and Lapata, 2020) uses content plans to generate synthetic datasets. *ConsistSum* (Ke et al., 2022) uses aspect and sentiment distribution to generate review-summary pairs. *MultimodalSum* (Im et al., 2021) generates summaries using multimodal data such as text, images, and meta-data. *TransSum* (Wang and Wan, 2021) uses aspect and sentiment embeddings to construct synthetic datasets. *COOP* (Iso et al., 2021) searches for convex combinations of latent vectors to generate summaries. *AceSum* (Amplayo et al., 2021) uses silver-labeled data obtained through seed words to train the model. *SW-LOO* (Shen et al., 2023) uses the aspect seed words to construct synthetic datasets, whereas *NLI-LOO* uses only aspects. *Acesum$_{ext}$*, *SW-LOO$_{ext}$*, and *NLI-LOO$_{ext}$* are the extractive versions respectively. *ASBOS* (Siledar et al., 2023b) uses aspect-sentiment to filter sentences and generate supervised summaries.

Multi-source Approaches. Due to the absence of any unsupervised approaches that use additional sources as input we fine-tune two models using our synthetic dataset for a fair comparison. BART-concat and T5-concat use BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) respectively with the input as a concatenated text. **Appendix G**.

6.4 Implementation Details

We used the `bart-large` (Lewis et al., 2019) and `t5-large` (Raffel et al., 2020) models from HuggingFace (Wolf et al., 2019). A learning rate of $2e-6$, batch size of 8, and 5 epochs performs the best on dev sets (Appendix H). During inference, we set beam size to 5 and no repeat ngram to 3. For encoding, we use the `all-MiniLM-L12-v2` from SBERT (Reimers and Gurevych, 2019). For SDC, $k = 8$ for Amazon and 10 for Oposum+ and Flipkart, whereas top-p selection is done using 85 percentile. Number of QA's are fixed to $M = 10$.

7 Results and Analysis

Automatic Evaluation. We use the ROUGE-{1,2,L} F1 score (Lin, 2004) (R1, R2 & RL) to assess the generated summary quality. Tables 3, 11

Gold (Amazon GPT-RDQ)
The **VuPoint Film and Slide Scanner** presents a mixed picture. It claims ease & efficiency for converting film to digital format, but user experiences differ. While some commend its user-friendliness and recommend it, others report frustrating issues like bleached images, compatibility problems, and subpar results. The need for XP compatibility limits its usefulness for modern systems. It offers potential for simple scanning but may require persistence to achieve desired outcomes.
MultimodalSum
I bought this product to scan my negatives. It does not work with Windows XP. I have tried to contact the company several times and have not received a response. I am very disappointed in the product. I would not recommend it to anyone.
BART-concat
I bought this Scanner to scan my slides and film. I have Windows XP and it does not work with it. I tried to download the drivers from the web site but they are not available. I called the company and they said they would send me a new one but I have not received it yet. I am very disappointed.
Our Model (MEDOS)
I purchased the **VuPoint FS-C1-VP Film and Slide Digital Converter** to scan my **35mm film** and slide negatives. It is not compatible with Windows XP. The software does not work with Windows 7 or 8. I have tried to contact the company and they do not respond to my emails. I would not recommend this product to anyone.

Table 5: Comparative Analysis. ChatGPT-generated summary using *reviews, description, and question-answers* (GPT-RDQ) followed by different model-generated summaries for an Amazon test set product. Information assisted from the description and question-answers are in **bold** and underline respectively. MEDOS is able to capture vital information from additional sources which won't be possible using only reviews.

& 12 present the results on Amazon and its variants, Oposum+ and its variants, and Flipkart and its variants respectively. In general, we observe that our MEDOS model performs better than baselines and outperforms MultimodalSum on all nine test sets. Better results on GPT-RDQ versions are expected as our model and these test sets use all sources for generating summaries. However, we observe that even on the original and GPT-R test sets our models perform much better. The reason for this we believe is that under the presence of multiple sources, our models are better at figuring out what information is essential and needs to be presented in the summary. Our approach to creating synthetic datasets plays a vital role in this. By showing the model the most relevant summary that takes into consideration all the sources, our models are able to learn better the task of opinion summarization as evidenced by the results. Next, almost for all cases, we observe that MEDOS performs better than the combination of simple concatenation approach and single encoder models (BART-concat & T5-concat). The MEDOS model due to its multi-encoder framework is able to selectively choose relevant information from the product description and question-answers. Additionally, we observe that single encoder models encounter context limitations in most cases thereby being unable to leverage the additional sources fully.

Qualitative Analysis. Table 4 presents the summary generated by our MEDOS model for four different products from the Amazon test set. Product description typically contains brand names as well as aspect-specifics. We observe that MEDOS excels at picking these specific names and including them in the generated summaries at appropriate places ensuring that the summaries are coherent. For example, **35mm film** in product 1 is an essential information that gets included in the summary. MEDOS also demonstrated the ability to pick relevant information from question-answers keeping the opinions being summarized in context. In product 4, the MEDOS model additionally gathers the compatibility of Nixon Coolpix L820 and the weight of the product from question-answers. Overall, MEDOS, due to its multi-encoder architecture and assistance from synthetic datasets during training learns to fuse relevant information well.

Comparative Analysis. Sample summaries generated by our model and some baselines on an Amazon test set product are shown in Table 5. MultimodalSum uses reviews, images, and meta-data, whereas Gold (Amazon GPT-RDQ), BART-concat, and our models use reviews, product description, and question-answers. In comparison to MultimodalSum, which also uses product description as part of the meta-data, MEDOS is able to capture details better such as **VuPoint FS-C1-VP Film and Slide Digital Converter** (brand name) and **35mm film** (information present only in description). In the presence of QA, MEDOS is able to provide relevant additional context to the information present in reviews. It picks details about Windows 7 and 8 from question-answers to present it along with the Windows XP. Finally, MEDOS does a better job compared to BART-concat in capturing details which we intuit is due to its multi-encoder framework. Additionally, the overall retention of the

	Amazon GPT-RDQ		
	R1 ↑	R2 ↑	RL ↑
MEDOS			
w. Reviews + Description + QA	**25.44**	**4.16**	**16.45**
w. Reviews + Description	23.54	2.43	14.81
w. Reviews + QA	20.05	1.36	12.90
w. Reviews	21.26	2.22	13.68

Table 6: Ablation study on Amazon GPT-RDQ. The highest utility comes from adding the description. QA in the presence of reviews and description aids the best.

consensus opinions from the reviews is unaffected.

Error Analysis. Unfortunately, our models are also prone to occasional hallucinations. For example, product 3 in Table 4 mentions that an adaptor is needed to charge iPhone 4s. Though, *needing an adaptor for some models* is mentioned in question-answers and *iPhone 4s* in reviews, there is no evidence of *iPhone 4s needing an adaptor*. We attribute such hallucinations to treating brand names such as *iPhone 4s*, *iPhone 5,* etc. as same.

Ablation Study. Table 6 presents the ablation study of our MEDOS model in using different sources on the Amazon GPT-RDQ test set. Results indicate that the combination of all sources performs the best. Intuitively, a higher score on Amazon GPT-RDQ summaries indicates that our model is leveraging the additional sources to generate more informative summaries. Without question-answers, we observe a 2 R1 point drop whereas, without the description a 5 R1 point drop. As expected, the utility of the description is higher than the question-answers. Descriptions contain aspect-specifics which help in enriching the summaries. In contrast, question-answers provide information related to specific queries about the product, which may or may not contribute to the overall summary. The distinction is evident, as using only reviews and question-answers results in poorer performance compared to using only reviews and description.

Human Evaluation. Table 7 shows the Best-Worst Scaling (Louviere et al., 2015) results, assessing the quality of opinion summaries. Six Masters' students aged 21-30 evaluated the model-generated summaries on: *faithfulness*, *coherence*, *conciseness*, and *fluency*. Each evaluator assigned a score of +1 for best, -1 for worst, and 0 for the remaining models. Final scores were computed by averaging the scores from all the evaluators. Notably, MEDOS achieved the best scores on all criteria.

SDC approach effectiveness. Our SDC approach

Amazon	Faithfulness ↑	Coherence ↑	Conciseness ↑	Fluency ↑
PlanSum	-0.50	-0.66	-0.63	-0.68
MultimodalSum	0.17	0.16	0.22	0.14
BART-concat	0.05	0.08	0.07	0.10
MEDOS	**0.21**	**0.41**	**0.23**	**0.50**

Table 7: Best-Worst Scaling. MEDOS generated summaries received better scores on all four criteria in human evaluation using the best-worst scaling method.

	Amazon GPT-RDQ		
	R1 ↑	R2 ↑	RL ↑
Our approach	**25.44**	**4.16**	**16.45**
Using only reviews for selection	21.36	2.04	13.86
Random selection	14.31	0.48	10.20

Table 8: SDC approach analysis. Our approach that uses description and question-answers along with reviews for selecting pseudo-summary performs the best.

selects the pseudo-summary based on description and QA first, followed by reviews. This ensures that the model sees relevant information during training thereby learning two things: picking of relevant information from additional sources and generating opinion summaries. Table 8 reports the results obtained using different SDC approaches.

Quantification of information captured. We measure the R1 scores of generated summaries with the sources on the Amazon test set to quantify the amount of information captured. Figure 2 shows our MEDOS generated summaries achieve an R1 of 18.64, 11.82, and 5.81 for reviews, description, and question-answers compared to 18.63, 8.28, and 5.46 for MultimodalSum. The nearly identical R1 for MEDOS and MultimodalSum suggest that even when additional information is present, MEDOS effectively captures all the crucial details from reviews. Next, MEDOS is better than both MultimodalSum and BART-concat in leveraging the information from description. Finally, for QA, R1 for MultimodalSum acts as a baseline as it does not use any QA during summarization. We observe that the BART-concat performs worse whereas MEDOS is able to capture relevant information.

MEDOS performance. We test the performance of MEDOS model by varying the number of parameters. Specifically, we use two variants of BART i.e. `bart-base` and `bart-large`, and report the results in Table 9. We observe that the `bart-base` variant of the MEDOS with just 0.3B parameters outperforms the single encoder models T5-concat and BART-concat (uses `bart-large`). In comparison

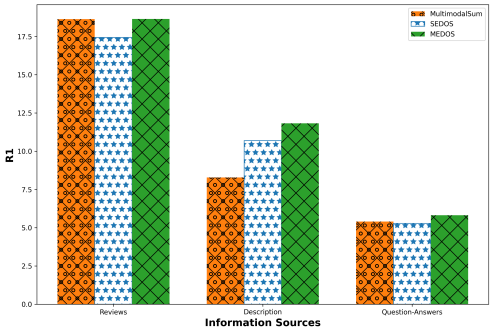

Figure 2: Quantification of information captured. MEDOS captures a similar amount of information from reviews as that of MultimodalSum, performs better for description, and picks relevant details from QA.

	mul?	#parameters	Amazon GPT-RDQ R1 ↑	R2 ↑	RL ↑
T5-concat	✗	0.7B	20.61	2.72	13.33
BART-concat	✗	0.4B	21.75	2.39	13.57
MEDOS					
bart-base	✓	0.3B	22.21	3.38	15.31
bart-large	✓	0.8B	**25.44**	**4.16**	**16.45**

Table 9: MEDOS Results. Comparison of MEDOS summaries for different parameter sizes. *mul?* represents models that use multiple encoders. #parameters indicate the number of parameters in billions (B).

between the two variants of MEDOS, we find that the `bart-large` version, as expected, performs better than `bart-base` due to a larger number of parameters. Overall, our findings indicate that the multi-encoder performs better and is able to capture details from different sources effectively.

LLMs on Multi-source Opinion Summarization. Recently, large language models (LLMs) have shown remarkable performance on a lot of tasks. For a fair comparison to baselines, we kept the focus of our work on smaller models in a self-supervised setting. For completion, we test the instruct models: Claude-2[3], Chatglm2-6b (Du et al., 2022), Llama-2-70b-chat[4], and Llama-2-7b-chat[5] (Touvron et al., 2023) on the task of multi-source opinion summarization. The training details of these models are not public and could possibly had access to test sets as a part of their training. We use the same GPT-RDQ prompts as in Appendix C

[3]https://www.anthropic.com/index/claude-2

[4]https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[5]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Model	#parameters	Amazon GPT-RDQ R1 ↑	R2 ↑	RL ↑
Claude-2	130B	31.11	4.73	16.67
Llama-2-70b-chat	70B	**32.77**	**7.84**	**20.28**
Chatglm2-6b	6B	27.31	4.72	16.80
Llama-2-7b-chat	7B	32.43	7.33	20.27
MEDOS	0.8B	25.44	4.16	16.45

Table 10: LLM results on Amazon GPT-RDQ test set compared to MEDOS.

to generate summaries using LLMs. We observe that our MEDOS model with just 0.8B parameters performs comparably to Claude-2 with 130B parameters and Chatglm-6b[6] with 6B parameters. Although Llama models with 70B and 7B parameters perform way better, for task-specific models MEDOS provides a cheaper alternative.

8 Conclusion and Future Work

We proposed a novel approach to create synthetic datasets by harnessing information from **reviews** and additional sources such as **product description** and **question-answers**. This method enables supervised training of models without the necessity of expensive annotated training datasets. Our proposed framework **MEDOS** uses separate encoders for selectively fusing information from these sources to generate an opinion summary. For evaluation, due to the absence of any test sets that contained such additional sources and annotated summaries, we extended the already available e-commerce test sets with additional sources and leveraged ChatGPT to annotate summaries. This resulted in six additional test sets with **662 opinion summaries** in total. Results show that our synthetic dataset approach and MEDOS framework outperforms the SOTA model on average by $\mathbf{14.5}\%$ and the simple input concatenation baseline by $\mathbf{6.5}\%$ across all nine test sets. Through qualitative and comparative analysis we demonstrated that our model-generated summaries are more informative and emphasize the importance of including additional sources for comprehensive summaries.

One future work is to expand these frameworks to encompass more reviews and all available sources, creating thorough product summaries.

[6]https://huggingface.co/THUDM/chatglm2-6b

Limitations

Our work, although uses a multi-encoder framework, is still currently limited by the size of the input. In e-commerce, reviews generally tend to be in the tens of thousands which could not be supported directly by the current model architectures. There has been research on increasing the context limits of the latest large language models, however, the performance of such models needs to be tested in the context of handling larger inputs for the task of opinion summarization. It becomes even more challenging to integrate additional sources found on product pages on e-commerce websites to provide an overall well-rounded product summary. Finally, we did not consider large language models (LLMs) in our work as our goal was to push for improvements in smaller models for multi-source opinion summarization utilizing only the available product corpus without the need for expensive large-scale annotated datasets and compute-intensive large-scale models. Our models do not use any LLM signals or LLM-generated data for training and rely only on the product corpus for learning the task of multi-source opinion summarization.

Ethical Considerations

We perform our experiments on existing opinion summarization datasets as well as extend the test sets by generating summaries using ChatGPT. Some of the examples in these datasets might not be appropriate for everyone. Our models may also propagate these unintended biases due to the nature of the datasets. We urge the research community to use our models and these test sets with caution and we are fully committed to removing any discrepancies in the existing datasets in the future.

References

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

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

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.

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).

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. 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*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Aspect-aware multimodal summarization for chinese e-commerce products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

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. *ArXiv*, abs/1908.08345.

Jordan J. Louviere, Terry N. Flynn, and Anthony A. J. Marley. 2015. Best-worst scaling: Theory, methods and applications.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Conference on Computational Natural Language Learning*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *ArXiv*, abs/1704.04368.

Ming Shen, Jie Ma, Shuai Wang, Yogarshi Vyas, Kalpit Dixit, Miguel Ballesteros, and Yassine Benajiba. 2023. Simple yet effective synthetic dataset construction for unsupervised opinion summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1898–1911, Dubrovnik, Croatia. Association for Computational Linguistics.

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023a. 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, Jigar Makwana, and Pushpak Bhattacharyya. 2023b. Aspect-sentiment-based opinion summarization using multiple information sources. *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing*.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

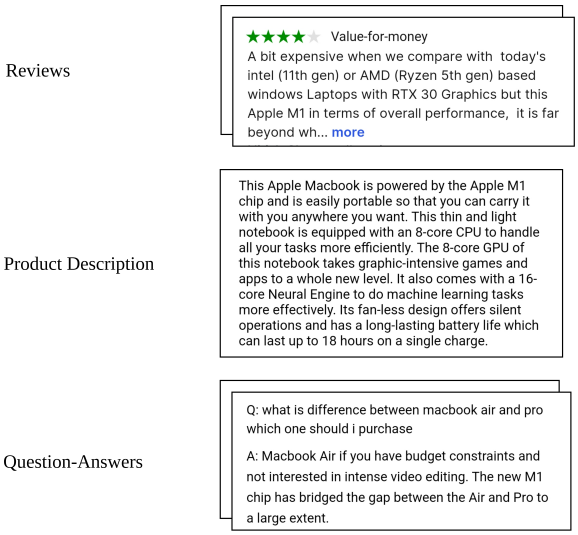

Figure 3: An example of *reviews*, *product description*, and *question-answers*. In our work, we consider multiple reviews and question-answers, and a single description per product to generate an opinion summary.

A Results on Oposum+ and Flipkart datasets

Results on Oposum+ and Flipkart and their corresponding extended test sets are reported in Tables 11 and 12 respectively.

B Information Sources

Figure 3 gives an example of the input sources. Per product, we consider multiple reviews and question-answer pairs and a single product description as the input for generating an opinion summary.

C GPT Prompts

GPT-R 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**: [r_1,...,r_k]*

GPT-RDQ prompt: *Following are the reviews, description, and question-answers 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, description, and question-answers only to generate the summary.*
***reviews**: [r_1,...,r_k]*

					Oposum+			Oposum+ GPT-R			Oposum+ GPT-RDQ		
abs?	**Model**	**R**	**D**	**Q**	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑
✗	Random	✓	✗	✗	33.63	10.79	19.82	24.08	2.38	13.25	23.68	2.12	12.98
✗	Oracle	✓	✗	✗	77.31	70.30	74.35	36.87	7.41	23.88	36.28	7.44	23.87
✗	QT	✓	✓	✓	37.72	14.65	21.69	25.82	3.47	14.01	25.81	3.21	14.13
✗	$\text{AceSum}_{\text{ext}}$	✓	✗	✗	38.48	15.17	22.82	-	-	-	-	-	-
✗	$\text{SW-LOO}_{\text{ext}}$	✓	✗	✗	40.45	19.13	23.20	-	-	-	-	-	-
✗	$\text{NLI-LOO}_{\text{ext}}$	✓	✗	✗	39.79	18.33	23.49	-	-	-	-	-	-
✓	CopyCat	✓	✗	✗	29.80	5.61	17.97	22.41	2.30	13.94	22.38	2.03	14.06
✓	AceSum	✓	✗	✗	32.98	10.72	20.27	22.78	3.59	13.20	23.54	3.51	13.88
✓	PlanSum	✓	✗	✗	30.26	5.29	17.48	22.37	2.05	13.32	22.64	2.25	13.71
✓	MultimodalSum	✓	✓	✗	33.08	7.46	19.75	23.35	2.98	14.53	23.73	2.80	14.70
✓	SW-LOO	✓	✗	✗	36.19	**12.17**	21.11	-	-	-	-	-	-
✓	NLI-LOO	✓	✗	✗	31.22	9.93	19.08	-	-	-	-	-	-
✓	T5-concat	✓	✓	✓	30.84	11.08	21.01	21.98	2.84	12.91	20.41	2.31	12.73
✓	BART-concat	✓	✓	✓	34.76	9.12	20.64	25.64	3.47	15.29	25.62	**3.36**	15.91
✓	**MEDOS**	✓	✓	✓	**36.57***	8.79*	**21.35***	**26.82***	**3.67***	**15.92***	**26.32***	3.34*	**16.10***

Table 11: Results on Oposum+ test set and its extensions. R, **D**, **Q** indicate the presence of *reviews*, *description*, and *question-answers* respectively in the input. *abs?* indicate abstractive systems. **Bold** and underline indicate best and second-best scores using abstractive systems. * indicates pvalue < 0.05 on paired t-test against MultimodalSum. Overall our combination of SDC approach and MEDOS model outperforms baselines across all three test sets.

					Flipkart			Flipkart GPT-R			Flipkart GPT-RDQ		
abs?	**Model**	**R**	**D**	**Q**	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑
✗	Random	✓	✗	✗	19.50	2.50	10.89	24.22	4.40	14.10	18.04	2.26	10.51
✗	Oracle	✓	✗	✗	34.07	6.34	21.30	38.35	9.98	24.81	29.47	5.12	19.20
✗	Clustroid	✓	✓	✓	21.42	3.01	12.08	27.76	5.56	16.77	10.17	1.45	7.74
✗	LexRank	✓	✓	✓	21.57	2.66	11.88	28.19	5.91	16.92	19.65	3.03	12.15
✗	QT	✓	✓	✓	25.18	3.62	13.05	30.94	5.96	15.34	22.92	2.95	11.97
✓	ASBOS†	✓	✓	✓	32.55	6.44	17.03	28.27	4.05	14.30	27.32	4.95	14.83
✓	CopyCat	✓	✗	✗	18.38	1.81	11.99	21.68	2.13	13.92	17.84	1.25	11.70
✓	PlanSum	✓	✗	✗	19.96	2.70	12.86	21.17	2.23	13.48	17.34	1.49	11.68
✓	MultimodalSum	✓	✓	✗	21.76	3.23	13.57	23.60	2.78	15.01	19.04	1.79	12.24
✓	T5-concat	✓	✓	✓	20.41	2.83	11.80	26.70	**5.75**	16.65	20.14	3.00	12.31
✓	BART-concat	✓	✓	✓	22.35	4.46	15.53	**27.27**	4.51	**17.22**	23.29	3.13	14.98
✓	**MEDOS**	✓	✓	✓	**25.97***	**5.29***	**16.05***	26.29*	4.03*	16.59*	**23.92***	**4.30***	**16.35***

Table 12: Results on Flipkart test set and its extensions. R, **D**, **Q** indicate the presence of *reviews*, *description*, and *question-answers* respectively in the input. *abs?* indicate abstractive systems. **Bold** and underline indicate best and second-best using abstractive systems. * indicates pvalue < 0.05 on paired t-test against MultimodalSum. † represents supervised systems. Overall our combination of SDC approach and MEDOS outperforms baselines.

description *: "..."*
question-answers*: [q_1,..,q_M]*

D Evaluation Metric

We use various metrics to qualitatively evaluate our model-generated summaries as well as ChatGPT-annotated summaries. We use the following:

1. **Informativeness**- how much of the information is captured?
2. **Faithfulness**- how consistent are the opinions compared to reference summaries?
3. **Coherence**- is the summary well organized and easy to read?
4. **Conciseness**- is the summary concise yet informative?
5. **Fluency**- is the summary fluent and grammatical?

E ChatGPT Annotation Quality

We assessed the GPT-generated summaries against human-written summaries on 5 metrics namely Informativeness, Faithfulness, Coherence, Conciseness, and Fluency. Results are presented in Table 15. We compare the ChatGPT-generated summaries against the human-annotated summaries for different test sets and report the results in Table

Rating	1	2	3	4	5
Informativeness	very poor	poor	acceptable	good	very good
Faithfulness	all hallucinated	somewhat verifiable	moderate hallucination	slight hallucination	no hallucination
Coherence	very poor	poor	acceptable	good	very good
Conciseness	verbose	moderately verbose	slightly verbose	almost concise	concise
Fluency	ungrammatical	slightly fluent	somewhat fluent	mostly fluent	fluent

Table 13: Human evaluation metrics. We use a scale of 1-5 to rate summaries on five evaluation metrics.

14. For ChatGPT-generated summaries refer to Table 19. GPT-R represents ChatGPT summaries using only reviews as input whereas GPT-RDQ represents ChatGPT summaries using reviews, description and question-answers.

		ChatGPT generated		
	No. of summaries	R1 ↑	R2 ↑	RL ↑
Amazon	96	25.09	2.58	14.02
Oposum+	90	30.01	4.42	15.30
Flipkart	145	30.20	4.18	15.74

Table 14: ChatGPT Results. Comparison of ChatGPT summaries with human-annotated summaries for different test sets.

	Info. ↑	Faith. ↑	Coh. ↑	Con. ↑	Flu. ↑
Human	3.88	3.91	3.68	3.83	3.62
GPT-R	4.02	4.13	4.02	4.09	3.98
GPT-RDQ	4.10	4.16	4.16	4.23	4.16

Table 15: Annotation quality. Both GPT-R and GPT-RDQ summaries score higher on all the metrics on average compared to human-annotated summaries. Scores range from 1-5. Info-*informativeness*, Faith-*faithfulness*, Coh-*coherence*, Con-*conciseness*, Flu-*fluency*.

F Dataset Details

Amazon Amazon contains reviews from 4 domains: *electronics, home & kitchen, personal care,* and *clothing, shoes & jewelry*. The evaluation set contains 3 summaries and 8 reviews per product. The training set contains ∼ 1M reviews over 90K products.

Oposum+ Oposum+ contains reviews from 6 domains: *bags, bluetooth headsets, boots, keyboards, televisions*. The evaluation set contains 3 extractive summaries and 10 reviews per product. The training set contains ∼ 4.13M reviews over 95K products.

Flipkart Flipkart contains reviews from 3 domains: *laptops, mobiles*, and *tablets*. The test set has 1 summary per product. The original test set contains 1K reviews per product on average. We downsample this to 10 reviews per product (randomly) for comparison.

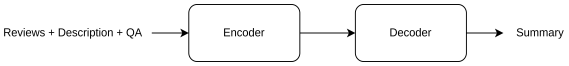

Figure 4: Framework of the baseline model that takes *reviews*, *description*, and *QA* as the input. A simple concatenation (+) of the input sources is used to generate a summary. During inference, the model generates a summary whereas during training the model uses pseudo-summary obtained through SDC process for learning.

G Single-Encoder Baseline

In the single-encoder framework, we concatenate reviews, product description, and question-answers using a separator symbol (</s>). This concatenated text c_{rdq} goes through an encoder to get the fused attention $\mathbf{a_f}$ as:

$$\mathbf{a_f} = \text{S-Attn}(\text{Enc}(c_{rdq})) \quad (12)$$

During training, the summary will be the pseudo-summary r and the input c_{rdq} will be formed using T, D, Q from the synthetic quadruplet. Figure 4 describes the single-encoder architecture. We use BART and T5 as our baseline models.

H Implementation Details

We used the Adam (Kingma and Ba, 2015) optimizer with eps of $1e-4$ and linear weight decay to optimize our models. We use learning rate in $[1e-6, 2e-6, 1e-5, 2e-5]$ and batch size in $[8, 16]$ as our hyperparameters. All experiments use NVIDIA A100-SXM4-80GB GPUs.

I Inter-Rater Reliability

We employed three professionals proficient in English in the age group of 23-34. Two evaluators were male and one was female. They were provided with detailed evaluation instructions along

with examples to rate summaries on different criteria as shown in Table 13. Each instance of the dataset was rated once and the work was equally divided among the three evaluators. 100 summaries were randomly chosen for evaluation and each evaluator annotated 50 summaries (25 unique and 25 common among all evaluators to compute Inter-Rater Reliability). Results of the evaluation can be found in Table 15. We first conducted a pilot study for evaluation with randomly sampled 10 summaries before proceeding to the final annotation. Table 16 shows the results of Fleiss' Kappa computed on different criteria.

	Human-annotated	GPT-R	GPT-RDQ
Informativeness	0.22	0.43	0.45
Factuality	0.24	0.36	0.44
Coherence	0.25	0.42	0.41
Conciseness	0.21	0.38	0.40
Fluency	0.24	0.45	0.41
Overall	0.23	0.41	0.42

Table 16: Fleiss' Kappa. We compute the Inter-Rater Reliability for human-annotated, GPT-R and GPT-RDQ on five metrics. GPT-R and GPT-RDQ scored higher on all the metrics compared to human summaries.

J SDC Approach Effectiveness

The novelty of our SDC approach lies in utilizing descriptions and question-answer pairs in the selection of pseudo-summaries in the most effective manner. The initial selection based on descriptions and question-answers ensures that the chosen pseudo-summary exhibits information overlap between these sources. This, in turn, aids the model in learning to extract information from these diverse inputs during the summarization process. Moreover, our strategy involves using the selected pseudo-summary to then identify the input reviews that are the most semantically close to it. This dual-step process enhances the model's learning of the opinion summarization task. Table 8 contains the results obtained using different SDC approaches. We find that our approach of creating synthetic datasets performs the best.

K MEDOS vs. LLMs?

Table 20 displays a comparison between the summaries generated by the LLM models and our MEDOS model. Our findings reveal that the MEDOS model adeptly captures most user opinions within the summary. However, LLMs go a step further, encompassing additional details to provide a comprehensive perspective on various product aspects. Despite this, our MEDOS model, significantly smaller and reliant solely on unsupervised corpus for synthetic datasets, competently extracts crucial user opinions without the extensive resources and fine-tuning required by LLMs, which often consist of billions of tokens and parameters.

Our primary goal was to leverage existing product data and refine smaller models like BART for multi-source opinion summarization, evaluating their effectiveness compared to ChatGPT. Prioritizing these smaller models aims to enhance accessibility and deployability, particularly on devices with limited resources. While LLMs outshine in performance, our focus on achieving high-quality outputs using smaller models within constraints represents a notable achievement. Insights gained from this endeavor can potentially enhance the data efficiency of larger models in the future. Beyond cost-effectiveness, MEDOS introduces a pathway to substantial results with reduced computational and data needs.

L Summary Lengths

Table 17 reports the mean summary length and mean standard deviations for summaries across three test sets: Amazon, Oposum+, and Flipkart.

	Amazon		Oposum+		Flipkart	
	μ	σ	μ	σ	μ	σ
Human annotated	55.20	12.98	82.16	20.54	118.86	37.11
GPT-R	58.31	13.01	89.61	8.90	82.71	13.54
GPT-RDQ	53.64	12.28	81.57	12.88	84.44	12.15
MultimodalSum	49.03	4.63	46.00	5.33	42.30	4.76
MEDOS	47.75	5.73	57.36	7.09	51.79	8.28

Table 17: Mean summary length (μ) and mean standard deviation (σ) for summaries corresponding to the three test sets: Amazon, Oposum+, and Flipkart.

M Example

Table 18 shows the *reviews, product description,* and *question-answers* for a sample product from the Amazon test set. Table 19 contains the human-annotated summaries from the original test set and our ChatGPT-generated (GPT-R and GPT-RDQ) summaries followed by different model-generated summaries for the same product.

Reviews
Exactly as described, at 8 + oz. of solid metal this grip offers a stable way to hold your lightweight digital camera, without putting you fingers in front of the lens or flash. I find it works well with the Kodak PlaySmart Video camera and the Nikon S9100 point and shoot. Opteka HG-5 Pistol Handgrip Stabilizer for Point-n-Shoot, DSLR and Video Cameras
Probably the best and the least expensive stick I've ever owned and I love it. I use this with my GoPro HD Hero2. It's a bit heavy but the construct is very good. You can use this as a weapon too. lol
Bought this as part of the stabilizer rig then realized that this was easier to use alone than the rig itself. I am going to use it with a camera with an active stabilizer. Videos looked good. Will update after I use it this weekend. It looks good and is built solid.
I use this with a dual-camera mount and I like this because of the heft / weight and it stays pretty secure whether I use it with the mount or on my flip video camera or snapshot. I'd recommend this handle highly.
I was planning to use this with my D7000 + Battery Grip + 80-200 f / 2.8 lens, but when I received it, I changed my mind. It just does not look like it can handle that load. I put it on my Panasonic GF2, and it performs very nicely. Would highly recommend it for lighter cameras.
The unit is quite sturdy. I bought it to replace the pistol grip unit also featured because the pistol grip locking mechanism did seem to want to lock tight. This unit locks in very tightly and also feels professional. A great purchase for the money.
This is the 2nd stabilizer that I've purchased one for my Sony a99 and one for my Sony a33. I can't speak highly enough about this handy little item! It's perfectly sized and the ergonomics is ideal! Two thumbs up!!
A low cost device that I bought and paired with a cell phone reduce jittery videos. Works pretty well for handheld use even when walking. The thread seemed a little recessed at first until I moved the washer flat. I recommend this product for anyone who records videos often for friends, and family especially with your cell phone.
Product Description
Opteka HG-1 Heavy-Duty Aluminum Ultra HandGrip Handheld Stabilization System for DSLR and Video Cameras. The Opteka HG-1 HandGrip Stabilization System is a video stabilization device designed specifically for point-and-shoots, Digital SLR cameras and compact camcorders. The Handgrip keeps your hands off the camera and allows you to capture videos from difficult angles. SpecificationsColor:Black; RedMaterials:Aluminum; Foam PaddedThread Size:1/4"Dimensions (HxLxW):6.25" x 1.5" x 1.5" (15.8cm x 3.8cm x 3.8cm)Weight:8.4oz (240g)
Question-Answers
What is on the bottom end? Is there a 1/4 - 20 female connection on the bottom? Yes it does have a 1/4 - 20 female connection very handy, i hope this helps you.
Does this work with nikon d800 It'll work with any camera that has a standard thread tripod socket. Note there is a male post at the top AND a female socket on the bottom. One of the handiest gadgets I've ever bought! If it only came in blue
Can this be used on a Nikon Coolpix L820, or is that camera too big / insufficient size? Yes you can. Use Can be used By any camera or camcorder Threaded for a tripod
Is the thread 1/4-20 Yes, 1/4 -20 (1/2 inch long) for standard tripod mount. threads right into the bottom of any small and midsized camera
If my arm shakes a lot, will this help? Probably not. I recommend you check out a mono pod or tripod. There also is a gimble style stabilizer that may help you but I've never used one so if you try it let me know how it works. Hope I answered your question.
I assume this can be screwed directly into a Canon VIXIA HF20? If your camera is threaded for a tripod it will work. It's a little heavy but it's handy to have
Does this handle water resistant and/or underwater? It's just a chunk of machined aluminum with a foam handle. Get it wet, it'll dry off. If it's screwed into an underwater housing you're golden, but if you just dunk your camera under water while attached to this attractive red and black handle, you'll have a nice expensive paperweight Comprende?

Table 18: Reviews, Product Description, and Question-Answers for a sample product from the Amazon test set.

Human-annotated summaries
Exactly as described. Looks good with a solid build and the unit is quite sturdy. This unit locks in very tightly and also feels professional. It is a great purchase for the money. I'd recommend this handle highly.
This stabilizer is perfect for light weight cameras and can also be used with a cell phone to help diminish jitter. You can feel the heft in this stabilizer and know that it will stay secure. It feels weighty and professional. I would say to use caution when using it with larger cameras but it is perfect for smaller cameras. Overall, it is recommended.
A great all-around, sturdy stick for cameras. It is very versatile and quite ideal for lightweight cameras especially, and this stick is highly recommended for its great price-point and overall versatility.
GPT-R summaries
The Opteka HG-5 Pistol Handgrip Stabilizer delivers on its promise. With a solid metal build, it secures lightweight cameras effectively. It impressed users across various devices, from Nikon to GoPro, proving its versatility. Sturdy and easy to handle, it's a budget-friendly choice for smoother videos. Whether used independently or as part of a rig, it offers stability and control. An ideal solution for steady shots, even when on the move.
The Opteka HG-5 Pistol Handgrip Stabilizer is a solid metal grip offering stability for lightweight cameras. Versatile with various camera types, it's praised for its heft, secure hold, and professional feel. Some found it surprisingly effective for reducing jittery videos, while others appreciated its durability and ease of use. Though not suitable for heavy setups, it's ideal for lighter cameras, making it a recommended, cost-effective choice for steady and smooth video capture.
Opteka HG-5 Pistol Handgrip Stabilizer is a solid metal grip offering stability for lightweight digital cameras like GoPro and point-and-shoots. It's versatile, working well both with rigs and standalone. Sturdy and secure, it handles various cameras effectively. A worthwhile purchase for smoother videos, whether shooting casually or on-the-go.
GPT-RDQ summaries
The Opteka HG-1 HandGrip Stabilization System is a versatile tool for capturing steady videos. Its sturdy build provides stability for lightweight digital cameras and camcorders. Users appreciate its ergonomic design and efficient angle adjustments. The grip's thread and weight offer ease of use, enhancing video quality even while on the move. An affordable solution for jitter-free recordings, especially useful for handheld shots and tricky angles.
Opteka HG-1 HandGrip offers solid stabilization for cameras. With its durable build and foam-padded handle, it adds stability to your shots. It accommodates various cameras, making it versatile for point-and-shoots, DSLRs, and compact camcorders. Itś especially handy for reducing jittery videos and capturing challenging angles. Its 1/4"-20 thread size ensures compatibility with most devices. The grip is praised for its heft and comfortable ergonomics. A cost-effective tool to improve video stability and angles for diverse cameras.
The Opteka HG-1 HandGrip Stabilization System proves its worth. Sturdy, versatile, and well-crafted, it stabilizes cameras for smooth video capture. Its ergonomic design offers a secure hold while preventing shaky shots. A valuable addition, especially for lightweight point-and-shoots and smaller cameras. Worth every penny for steady, shake-free videos.
PlanSum
i bought this for my nikon d7000 and it works great. i have had no problems with it. the only thing i don't like is that it is a bit bulky, but it's not a big deal for me. if you are looking for a good camera, this is the best way to go.
MultimodalSum
This is a great product. I use it with my Canon T3i and it works great. It is a little heavy, but that is to be expected. I like the fact that it has a solid feel to it. I would recommend this product to anyone.
BART-concat
I bought this to use with the Opteka HG-5 Pistol Handgrip Stabilizer for my Canon Powershot T3i and it works great. It is a bit heavy, but that is to be expected with a tripod.
Our Model (MEDOS)
The Opteka HG-1 Heavy-Duty Aluminum Ultra HandGrip Handheld Stabilization System for DSLR and Video Cameras is a great product. I use it with my Nikon Coolpix L820 and it works great. It is a little heavy, but that is to be expected for a small camera.

Table 19: Three human-annotated summaries, three GPT-R summaries, there GPT-RDQ summaries followed by the summaries from different models for the same product as in Table 18.

Gold (Amazon GPT-RDQ)
The VuPoint Film and Slide Scanner presents a mixed picture. It claims ease & efficiency for converting film to digital format, but user experiences differ. While some commend its user-friendliness and recommend it, others report frustrating issues like bleached images, compatibility problems, and subpar results. The need for XP compatibility limits its usefulness for modern systems. It offers potential for simple scanning but may require persistence to achieve desired outcomes.
Chatglm2-6b
scanner from VuPoint. The scanner has worked great for me and I have been able to scan over 1, 200 35mm negatives and a dozen slides. I have also been able to upload a few black and white negatives from 1963. The scanner is user-friendly and efficient tool. The software that comes with the scanner is good for Windows XP and Vista, but the company prefers to sell other products. I am looking for a scanner that can work with more modern versions of Windows, but I am not satisfied with the VuPoint company. I am also waiting for a replacement unit to be delivered.
Claude-2
Reviews of this scanner are quite mixed. Some find it easy to use and good for digitizing old negatives and slides. But many cite issues getting it to work properly, with several noting it is not compatible with Windows versions after XP. Complaints include poor image quality, too much light bleaching images out, and tech support being expensive. The included software also gets some negative feedback for lacking updates. A few reviewers say the holders are fiddly for fitting negatives into. Overall while some find it efficient for converting film, many note disappointing quality and outdated software/hardware compatibility as negatives.
Llama-2-7b-chat
The VuPoint FS-C1-VP Film and Slide Digital Converter has received mixed reviews. Some users have had issues with image quality, light bleaching, and poor customer support, while others have found the scanner easy to use and effective at converting slides and negatives to digital format. The product's compatibility with newer operating systems and software updates has also been a concern for some reviewers. Overall, the product's performance and customer service have been a disappointment for some buyers.
Llama-2-70b-chat
The VuPoint FS-C1-VP Film and Slide Digital Converter has mixed reviews. Some reviewers find it easy to use and efficient for scanning 35mm film negatives and slides, while others have issues with image quality, software compatibility, and customer support. One reviewer had a positive experience with the company's customer service, while another found it unhelpful. The product's ability to scan images at 5.0 Mega pixel quality and automatic color balance and exposure control are mentioned in the description, but some reviewers still express dissatisfaction with the product's overall performance.
Our Model (MEDOS)
I purchased the VuPoint FS-C1-VP Film and Slide Digital Converter to scan my 35mm film and slide negatives. It is not compatible with Windows XP. The software does not work with Windows 7 or 8. I have tried to contact the company and they do not respond to my emails. I would not recommend this product to anyone.

Table 20: Comparative analysis with LLM generated summaries. ChatGPT-generated summary using *reviews, description, and question-answers* (GPT-RDQ) followed by different LLM-generated summaries and our MEDOS model generated-summary for an Amazon test set product.