EvioSum: An Evidence-Guided Generation Framework for Faithful and Interpretable Opinion Summarization

Jian Wang
Shandong University
Jinan, China
wangjian026@126.com

Yanjie Liang
Shandong University
Jinan, China
202335329@mail.sdu.edu.cn

Yuqing Sun*
Shandong University
Jinan, China
sun_yuqing@sdu.edu.cn

Bin Gong
Shandong University
Jinan, China
gb@sdu.edu.cn

Abstract

The faithful and interpretable opinion summarization aims to generate a summary that captures the diverse opinions expressed in a document set while providing explanations for the divergences between these opinions. In this paper, we propose an evidence-guided framework to enhance opinion coverage and provide divergence explanations. It first generates the majority opinion as an initial summary and partitions the source documents into multiple evidence sets based on their relevance to the majority opinion. Then, a summary extension strategy is employed to expand the initial summary by incorporating different opinions from these sets. The framework also employs a submodular optimization algorithm to select evidence from different evidence sets in order to reflect the divergences between opinions. Experiments on two benchmark datasets demonstrate that our method outperforms multiple baselines in terms of both the lexical and semantic consistency with reference summaries, while having low computational overhead. Ablation studies confirm that both the document partition and summary extension mechanisms contribute to the model performance. The LLM-based and human evaluation results also show that our method can identify more comprehensive evidence that better captures opinion divergences.[1]

CCS Concepts

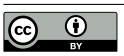
• **Computing methodologies** → **Natural language generation**.

Keywords

Faithful opinion summarization; Interpretable; Evidence; Aspects

ACM Reference Format:
Jian Wang, Yanjie Liang, Yuqing Sun, and Bin Gong. 2026. EvioSum: An Evidence-Guided Generation Framework for Faithful and Interpretable Opinion Summarization. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26), February 22–26, 2026, Boise, ID, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3773966.3777962

*Corresponding author.
[1]The source codes are accessible by the link: https://github.com/wangjian026/EvioSum



WSDM '26, Boise, ID, USA

ACM ISBN 979-8-4007-2292-9/2026/02
https://doi.org/10.1145/3773966.3777962

1 Introduction

For a given topic, a large number of opinions are shared on the web. These opinions vary in emotional orientation, and factual interpretation. To efficiently capture both the commonalities and divergent focuses among them, the task of faithful and interpretable opinion summarization aims to generate summaries that cover the diverse opinions in source documents while providing explanations for the opinion divergences [4, 32]. This task has important applications in areas such as the public policy and social issue research. However, the inherently subjective nature of opinions makes summarization challenging and difficult to interpret.

Large language models (LLMs) have become the main choice for the summarization task due to their strong capabilities in text comprehension and generation [10, 20, 32]. However, they tend to focus on the main points in documents based on relevance, which often causes the generated summaries overlook some minority opinions [12, 18, 38]. For this problem, the prompt-engineering based methods such as CPSum[32] and SummIT [35] iteratively construct fine-grained prompts to enhance the focus of LLMs on minority opinions based on the summary evaluation results. The evaluation process is usually repeated several times, resulting in a large computational overhead. The chain of thought (COT)[16] based methods TCG [3] and AspDeSum [21] first cluster the documents based on aspects, then generate sub-summaries for each cluster, and finally merge the sub-summaries into a complete summary. These methods improve the opinion coverage of the summaries with respect to the source documents but without any explanations of the summaries. Furthermore, these methods focus on product review data with relatively fixed aspects and are difficult to adapt to more free-form social content such as tweets or blogs.

For the above challenges, we propose the **Evi**dence-guided **o**pinion **Sum**marization framework (**EvioSum**). To enhance opinion coverage, the framework first guides LLMs to generate the majority opinion. Based on the relationship between each document and the majority opinion, the source documents are then partitioned into a support evidence set, a divergent evidence set, and a neutral evidence set. Then the framework extends the majority opinion to form the summary by incorporating the divergent opinion derived from the divergent evidence set, followed by the neutral opinion

summarized from the neutral evidence set. This summarization process establishes the association between each opinion in summary and its corresponding evidence, thereby improving the summary's opinion coverage and interpretability. To reflect the divergence between the majority and divergent opinions, EvioSum constructs an *explanation set* of evidence pairs by selecting evidence from their respective sets. Specifically, it introduces an aspect-enhanced evaluation method to assess each candidate *explanation set* based on three criteria: support for the corresponding opinion, divergences between the paired evidence, and overall content completeness and distinctiveness. Then the selection process is formulated as a submodular optimization problem, which allows for an efficient approximate solution using a greedy algorithm.

We conduct experiments on two benchmark datasets, and the results show that our method outperforms the state-of-the-art (SOTA) methods across multiple metrics and human evaluation, while incurring lower computational overhead. Ablation studies confirm the effectiveness of both the document partition and summary extension components. We evaluate the quality of the *explanation set* using a high-performance LLM and human evaluation. The results show that, compared with two baseline methods, our approach identifies more comprehensive evidence that better reflects the divergences between opinions.

2 Related works

Due to the computational cost of constructing high-quality supervised data, existing researches have primarily focused on unsupervised methods. These methods can be broadly categorized into two groups. One is the task-tuned methods that aim to inject the domain-specific knowledge into the model. Another type is the LLM-based methods that directly leverage the knowledge embedded in LLMs. The following sections provide a detailed discussion of these methods. Additionally, we summarize several insightful interpretability techniques relevant to the summarization task.

2.1 Task-tuned methods for summarization

These methods primarily fall into two categories: those that construct pseudo-summaries based on key information selection, and those that leverage autoencoder frameworks for self-supervised training based on the domain data.

To construct the pseudo-summaries, the ConsistSum method [14] defines the distance between two documents based on sentiment and aspect, and selects the document that is most central among the source documents as a pseudo-summary. Similar to ConsistSum, OPINESUM slices the source documents into multiple propositions and selects those that are widely entailed by the other documents to construct the pseudo-summary [26]. For the second category, COPYCAT [5] trains a variational autoencoder (VAE) [15] by reconstructing a source document from the remaining documents, with the innovativeness of the reconstruction controlled via latent variables. At inference time, the model constrains novelty to be minimal, encouraging the generation of consensus opinions. MeanSum[7] learn the representation of each document by reconstructing the input documents[7], and generates a summary by decoding from the average semantic representation of multiple input documents. While both approaches aim to model domain-relevant content, they often underrepresent minority opinions, limiting their effectiveness for faithful opinion summarization.

2.2 LLM-based method for summarization

These methods primarily focus on designing prompts or decomposing the summarization task into multiple simpler sub-tasks that LLMs can handle, thereby effectively leverage the knowledge in LLM.

To construct high-quality prompts, the CPSum, Summit and Self-Refine methods use LLMs to evaluate the generated summaries and calibrate the prompts based on the evaluation results[27, 32, 35]. For instance, CPSum iteratively selects high-quality opinions from the generated summaries, and incorporates them to the prompts for guiding LLMs to retain important opinions and explore new ones [32]. While this kind of methods enhances the focus of LLMs to minority opinions, the iterative process incurs high computational overhead. Inspired by the chain-of-thought (CoT) technique [17], several methods decompose the opinion summarization task into a series of subtasks and design tailored prompts for each subtask, thereby reducing the overall task complexity [3, 20, 21, 34]. For example, AspDeSum [21] adopts LLMs to cluster the documents based on aspects, generates a aspect-level summary for each cluster, and then merges all aspect-level summaries into a complete summary. Another representative method, SentiSum [20], adopts a three-layer sentiment consolidation framework that emulates the human meta-review writing process, where each layer accomplishes a distinct summarization-related subtask by an LLM. The above methods enhance the semantic coverage of summaries with respect to the source documents. However, they do not consider the complementary or conflicting relationships between different opinions throughout the summarization process, leading to a lack of logical coherence among the opinions presented in the summaries.

2.3 Interpretable techniques for summarization

The tradition interpretability techniques primarily focus on supervised methods, leveraging gradients and attention weights for model interpretability. Gradient-based interpretation methods were originally developed in the image domain, where the saliency score (SS) of a pixel is computed by calculating the gradient of the model output with respect to that pixel [30]. In the context of summarization models, the saliency of each input word can be obtained by computing the Euclidean norm of the gradient with respect to its word embedding [13]. Attention-based interpretation methods assume that attention weights are positively correlated with the importance of the corresponding input positions [2, 11, 33]. Thus, they achieve interpretability through outputting attention weights. However, generating opinion summaries requires logic reasoning across multiple documents, making it challenging to express interpretability merely by identifying a few key words. Recently, the representative methods such as ESCA[31], RTSUM[6] and SP[28] introduce intermediate structures or control units into the summarization process to enhance the interpretability of the generated summaries. The quality of the explanations provided by these methods largely depends on the guidance of supervised data. In the absence of such supervision, the explanations they produce often lack accuracy and generalizability.

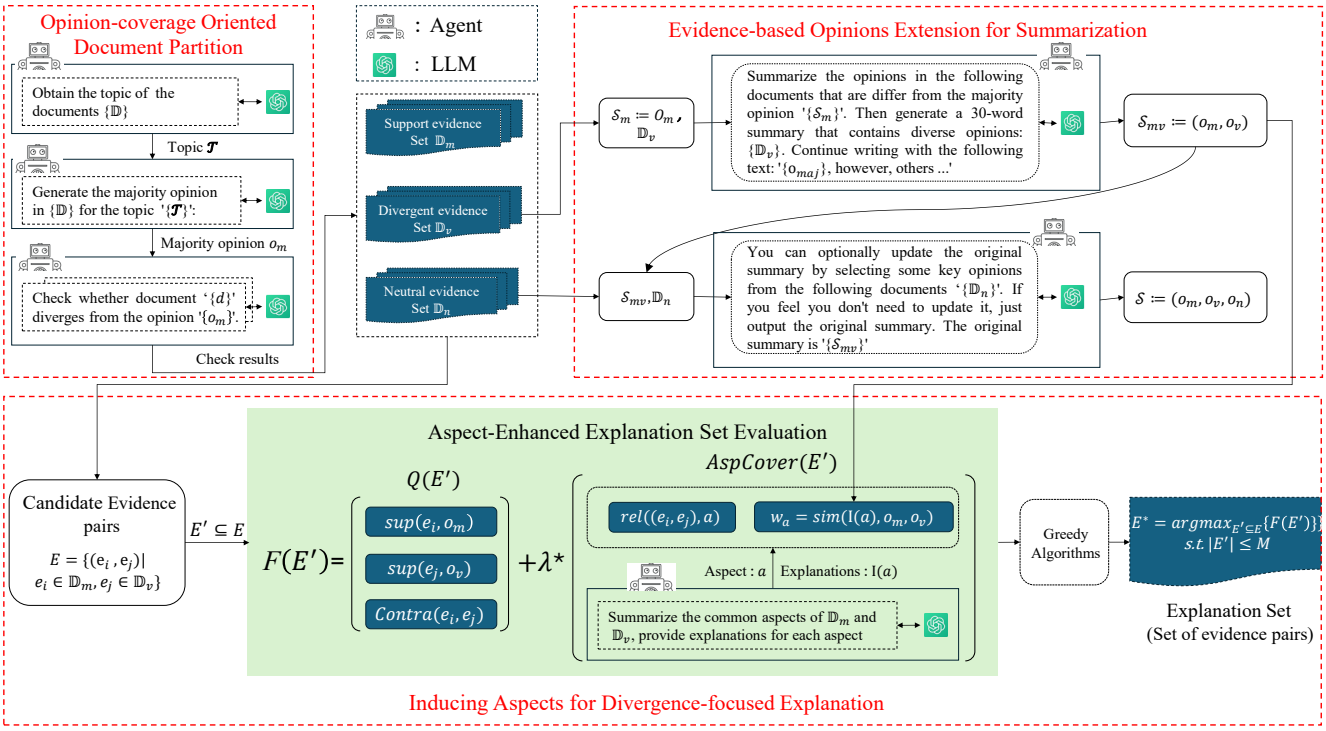


Figure 1: The evidence-guided framework for faithful and interpretable opinion summarization. The agent in this framework refers to an autonomous module that applies the designed prompts to guide LLMs.

3 Problem and framework

Given a document set $\mathbb{D} = \{d_1, d_2, \cdots, d_{|\mathbb{D}|}\}$, the faithful and interpretable opinion summarization aims to generate a summary S that covers the diverse opinions in $\mathbb{D}$ while providing explanations for the divergences between these opinions. For this task, we propose an evidence-guided opinion summarization framework (**EvioSum**), which is illustrated in Fig. 1 and the details are given below.

Our framework includes three components. The first component generates the majority opinion and partitions the document set $\mathbb{D}$ into three subsets, namely the support evidence set $\mathbb{D}_m$, divergent evidence set $\mathbb{D}_v$ and neutral evidence set $\mathbb{D}_n$, respectively. The second component takes the majority opinion as the initial summary and incrementally extends it by incorporating the divergent opinions o_v from $\mathbb{D}_v$ and the neutral opinions from $\mathbb{D}_n$. The third component selects representative evidence from $\mathbb{D}_m$ and $\mathbb{D}_v$ to construct an *explanation set* for explaining the key divergences between o_m and o_v.

3.1 Opinion-coverage Oriented Document Partition

To ensure the opinion coverage of the summary, we partition the source documents based on the majority opinion, as it is the most salient opinion in the documents and can serve as a reference opinion for interpreting and contrasting other opinions. To get the majority opinion, we first extract the topic $\mathcal{T}$ of $\mathbb{D}$ by an LLM, denoted as $\mathcal{T} = LLM(\mathbb{D})$. Then the majority opinion o_m is obtained by prompting LLM under $\mathcal{T}$: $o_m = LLM(\mathbb{D}, \mathcal{T})$. Guided by the majority opinion, the document set $\mathbb{D}$ is partitioned based on the relationships between each document $d \in \mathbb{D}$ and the majority opinion. To better capture the differences between opinions, rather than a simple opposition, we restrict the relationship types to three categories: **support** for the majority opinion, **divergence** from the majority opinion, and **neutrality** to the majority opinion. We use function $\mathbb{I}(d, o_m) \in \{1, -1, 0\}$ to denote the relationship between document d and opinion o_m obtained by an LLM, where values 1,-1,0 denotes the *support*, *divergence*, and *neutrality* relationships, respectively. As a result, $\mathbb{D}$ is partitioned into the support evidence set $\mathbb{D}_m = \{d \in \mathbb{D} | \mathbb{I}(d, o_m) = 1\}$, divergent evidence set $\mathbb{D}_v = \{d \in \mathbb{D} | \mathbb{I}(d, o_m) = -1\}$ and neutral evidence set $\mathbb{D}_n$.

3.2 Evidence-guided Opinion Extension for Summarization

To generate a faithful opinion summary, we adopt a step-wise generation process that begins with the majority opinion and progressively incorporates other opinions derived from various evidence sets. Specifically, we first initialize the summary with the majority opinion, denoted as $\mathcal{S}_m := (o_m)$. Then, for the divergent evidence set $\mathbb{D}_v$, we guide an LLM to generate an extended summary by combining the initial opinion o_m with the new opinions from $\mathbb{D}_v$, denoted as $\mathcal{S}_{mv} = LLM(o_m, \mathbb{D}_v) := (o_m, o_v)$, where o_v is the new extended opinion by $\mathbb{D}_v$, we call o_v as the divergent opinions. For

the neutral evidence set $\mathbb{D}_n$, the content is either irrelevant to the topic or has different perspectives from the majority and divergent opinions, thus we use it to supplement $\mathcal{S}_{mv}$, both to refine the semantics of $\mathcal{S}_{mv}$ or to extract other opinions that are differ from $\mathcal{S}_{mv}$, formally: $\mathcal{S} = LLM(\mathcal{S}_{mv}, \mathbb{D}_n) := (o_m, o_v, o_n)$, where o_n represents the newly interpreted opinions from $\mathbb{D}_n$.

3.3 Inducing Aspects for Divergence-focused Explanation

Given two opinions o_m and o_v with their respective evidence sets $\mathbb{D}_m$ and $\mathbb{D}_v$, this component aims to construct an *explanation set* E^* that contains multiple evidence pairs:

$$E^* = \arg\max_{E' \subseteq E} F(E') \quad \text{s.t.} \quad |E'| \leq M \tag{1}$$

where $E = \{(e_i, e_j) \mid e_i \in \mathbb{D}_m, e_j \in \mathbb{D}_v\}$, and M denotes the maximum allowed number of evidence pairs in E^*, $F()$ denotes an function that evaluates the set from the following three dimensions:

(1) High Support: e_i and e_j should support their corresponding opinions o_m and o_v, respectively.
(2) High Divergence: e_i and e_j in each pair exhibits contradiction in at least one shared aspect. An aspect refers to a specific argumentative dimension of opinion.
(3) Distinctiveness and Completeness: Evidence pairs should be distinct from one another and collectively cover multiple aspects.

To construct E^*, we first define the function $F()$, and then use a submodular optimization algorithm to select evidence pairs that maximize $F(E^*)$.

3.3.1 Evaluation of Explanation set. For the first two evaluation dimensions, let $sup()$ denote a function that returns the support score, indicating how strongly the evidence supports the given opinion. Let $contra()$ denote another function to return the degree of divergence between two pieces of evidence. $sup()$ and $contra()$ can be implemented using an LLM or a pre-trained natural language inference (NLI) model. Here, considering the computational cost, we employ an NLI model and use the entailment probability predicted by the model to represent the degree of support, while the contradiction probability represents the degree of divergence. Then for a candidate set $E^{'} \subseteq E$ of evidence pairs, the evaluation results on the two dimensions are represented as follows, where α is a hyperparameter:

$$Q(E^{'}) = \sum_{(e_i, e_j) \in E^{'}} \Big[\alpha \cdot sup(e_i, o_m) \cdot sup(e_j, o_v) + (1 - \alpha) \cdot contra(e_i, e_j)\Big] \tag{2}$$

To measure the distinctiveness and completeness of the evidence pairs contained in $E^{'}$, we first induce the aspects covered by the evidence. Let $\mathcal{A}$ be the set of aspects that appear in both $\mathbb{D}_m$ and $\mathbb{D}_v$. We employ an LLM to extract $\mathcal{A}$ along with a textual description $I(a)$ for each aspect $a \in \mathcal{A}$:

$$(a, I(a)) = \text{LLM}(\mathbb{D}_m, \mathbb{D}_v) \tag{3}$$

Since different aspects should have different weights. The higher the relevance of an aspect to the opinions o_m and o_v, the better it can capture the core divergences. To represent this property, we quantify the weight of each aspect by calculating the semantic similarity between its aspect description and the two opinions:

$$w_a = \frac{sim\,(I(a), o_m) + sim\,(I(a), o_v)}{2} \tag{4}$$

where, $sim(\cdot, \cdot)$ denotes semantic similarity between two texts. Here we choose the cosine similarity between the embeddings of the two texts as $sim(\cdot, \cdot)$.

Next, we quantify the relevance between an evidence pair and an aspect so as to compute the completeness of all evidence pairs with respect to the aspect set $\mathcal{A}$. Specifically, for an evidence pair $p = (e_i, e_j) \in E^{'}$, its relevance to an aspect $a \in \mathcal{A}$ is defined as:

$$rel(p, a) = sim(e_i, I(a)) * sim(e_j, I(a)) \tag{5}$$

and the completeness $AspCover(E^{'})$ of the *explanation set* is defined as the degree to which all aspects are represented in the set:

$$AspCover(E^{'}) = \sum_{a \in \mathcal{A}} w_a \sqrt{\sum_{p \in E^{'}} rel(p, a)} \tag{6}$$

where, the square root is applied to reflect the diminishing returns: as more evidence pairs cover the same aspect, the incremental contribution decreases, preventing redundancy of any single aspect.

Finally, we combine the above three dimensions to form the final evaluation function:

$$F(E^{'}) = Q(E^{'}) + \lambda \cdot \text{AspCover}(E^{'}) \tag{7}$$

Here, λ is a hyperparameter that balances the two items based on the evaluation priorities and can be flexibly adjusted according to the application scenario.

3.3.2 Greedy Evidence Pair Selection. We treat the construction of *explanation set* E^* as a submodular optimization problem, i.e., maximizing $F(E^*)$ under the constraint $|E^*| \leq M$. A submodular function is a set function with the diminishing returns property, meaning the marginal gain of adding an element to a set decreases as the set grows[23]. The $F()$ is a non-decreasing concave function that satisfies the properties of a submodular function, allowing the optimization problem to be effectively solved using a greedy algorithm. We adopt the greedy algorithm 1 to construct E^*. To avoid redundancy among evidence pairs, we stop collecting new pairs when the marginal gain of adding a pair falls below a predefined stop proportion of the gain obtained in the previous step.

4 Experiments

4.1 Datasets and Metrics

We use the Microblog Opinion Summarization datasets (MOS) [4] for experiments. MOS contains two sub-datasets with different topics: the UK Election Opinionated Dataset (**EO**) and the COVID-19 Opinionated Dataset (**CO**). These two datasets are collected from the Twitter site[2], where **EO** contains documents related to the election topics, and **CO** focuses on COVID-19 related topics. Each sample in the two datasets includes multiple documents and a manually generated summary that contains one or more opinions. We evaluate the generated summaries using multiple metrics from different perspectives. To assess the content similarity between generated summaries and human-written references, we employ

[2]https://twitter.com

Algorithm 1: Algorithm for constructing *explanation set*

Input: Evidence sets $\mathbb{D}_m$, $\mathbb{D}_v$, maximum number of evidence pairs M in E^*
Output: The *explanation set* E^*

Initialize $E^* \leftarrow \emptyset$, $E \leftarrow \{(e_i, e_j) \mid e_i \in \mathbb{D}_m, e_j \in \mathbb{D}_v\}$, stop proportion γ, temporary marginal gain $\Delta_{\text{prev}} = 0$;
while $|E^*| <= M$ **do**
  $(e_i^*, e_j^*) \leftarrow \arg\max_{(e_i,e_j)\in E\setminus E^*} \{F(E^* \cup \{(e_i, e_j)\}) - F(E^*)\}$;
  $\Delta(e_i^*, e_j^*) \leftarrow F(E^* \cup \{(e_i^*, e_j^*)\}) - F(E^*)$; // The marginal gain
  if $\Delta(e_i^*, e_j^*) >= \gamma \cdot \Delta_{prev}$ **then**
    $E^* \leftarrow E^* \cup (e_i^*, e_j^*)$;
    $\Delta_{\text{prev}} \leftarrow \Delta(e_i^*, e_j^*)$;
  end
end
return E^*

ROUGE [22] and BERTScore [36]. For the opinion coverage evaluation, we adopt GEVAL [24], a metric that uses LLMs to perform generative scoring by designing coverage-focused prompts.

4.2 Implementation Details

We choose Vicuna-7b as the backbone model for our framework. To preserve its exploratory capacity, we set the temperature to 0.7. We preprocess the source documents by removing usernames and invalid character codes that do not contribute to content understanding. For achieving the similarity function $sim()$, we use the all-MiniLM-L6-v2 model provided by the sentence-transformers library[3] to encode the input texts and compute the cosine similarity between their embeddings. We set the α to 0.5, λ to 2 and γ to 0.3. The maximum number of evidence pairs M is set to 3. We use the *roberta-large-mnli*[25] as the NLI model. For ROUGE evaluation, we use the following configuration via PyRouge, a Python wrapper for the ROUGE toolkit: ROUGE-1.5.5.pl -f A -a -c 95 -m -n 2 -2 4 -u -p 0.5. Considering the invocation overhead, we use GPT-3.5-trubo as the base LLM for GEVAL. The experiments are conducted on V100 GPUs. For our experimental results, we repeat the experiment 4 times and take the mean value. All the prompts used in our method are shown in Fig. 3.

4.3 Comparison Methods

We compare our method with the following methods: **Claude-3.5-haiku**, **Claude-3.5-haiku**,[4] **GPT-4o-mini** [5] and **DeepSeek-V3** [6] are the commonly used LLMs on multiple tasks. We use the following prompts for guiding these LLMs: *'Summarize the following documents to generate a summary that contains the majority and other opinions. The given documents are {𝔻}. If there is no other opinions, only output the majority opinion.'* **LexRank** adopts the *PageRank* algorithm to select important sentences based on a weighted graph [8]. **QT** clusters similar sentences, quantifies the popularity of each cluster, and extracts sentences from the most popular one [1]. **SummPip** aggregates sentences into multiple clusters, and compresses each cluster to construct a summary [37]. **Copycat** models the summarization process with a hierarchical variational autoencoder [15] and uses a pointer-generator mechanism [29] to generate summaries [5]. **OPINESUM** [26] constructs pseudo-summaries by a textual entailment model and uses these summaries to train a summarization model. **CPSum** [32] is the SOTA method on the MOS datasets. It iteratively optimizes prompts through the feedback of the generated summary. **BART-base** [19] serves as a supervised baseline, trained on the full training sets of both the EO and CO datasets. **EvioSum** is our method.

4.4 Main Results

The main comparison results are shown in Table 1. The results marked with '*' are taken from the paper [4]. R-1, R-2, R-L and R-SU4 denote ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-SU4 [22], respectively. BS stands for BERTScore [36]. IC stands for the complexity of LLM invocation. The results of CPSum come from the original paper. The first block in Table 1 contains the results of LLMs, and the second block includes unsupervised methods specifically designed for opinion summarization tasks. We also show the results of supervised BART-base method as the reference. The results show that EvioSum outperforms all baseline methods across multiple metrics. Compared to the state-of-the-art (SOTA) method CPSum, EvioSum achieves better results for most evaluation metrics while incurring lower overhead of LLM invocation. Specifically, CPSum uses LLMs to check the support relationship between each source document and each sentence in the generated summary, and uses the check results to construct feedback for calibrating prompts. This process is repeated several times until a predetermined stopping condition is reached. As a result, the complexity of LLM invocation is $O(|\mathbb{D}|^2)$, whereas the complexity of our method is $O(|\mathbb{D}|)$.

Comparing the results of LLMs from the view of model size, the performance improvements do not consistently align with increases in model scale. Based on this observation, we argue that, rather than simply employing larger models, developing effective opinion-guidance strategies for steering LLMs plays a more critical role in enhancing summarization quality. We also observe that the supervised model BART-base outperform existing LLM-based methods in terms of ROUGE metrics. However, for the semantic similarity-based metric BERTScore, supervised methods do not exhibit strong performance. This indicates that the BART-base model is more dependent on lexical overlap, whereas LLM-based methods are better at capturing semantics.

5 Model Analysis

5.1 Evaluation of Opinion Coverage

The opinion coverage can be evaluated in two perspectives, one is against the source documents and another is against the reference summary. High opinion coverage reflects that the key opinions in source documents or reference are adequately covered by the summary. However, summaries with high coverage may include repeated or overlapping content, reducing its clarity. To address the above issue, we incorporate redundancy as an additional factor

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[4] https://www.anthropic.com/claude/
[5] https://openai.com/gpt-4
[6] https://www.deepseek.com/

Table 1: Model Comparison Results

Method	Election Opinionated Data(EO)					CoVID-19 Opinionated Data(CO)					
	R-1	R-2	R-L	R-SU4	BS	R-1	R-2	R-L	R-SU4	BS	IC
LLMs:											
Claude-3.5-haiku	29.56	9.22	25.28	10.83	0.859	26.32	8.30	22.90	10.02	0.850	O(1)
Claude-3.5-sonnet	28.90	9.10	25.14	10.53	0.858	25.67	7.80	22.43	9.85	0.850	O(1)
DeepSeek-V3	29.80	8.91	25.91	10.63	0.857	25.78	7.53	22.74	9.33	0.848	O(1)
GPT-4o-mini	28.44	8.20	24.35	10.51	0.857	25.05	6.95	21.67	9.25	0.849	O(1)
Unsupervised methods:											
LexRank*	14.27	1.15	9.62	–	0.856	16.41	1.48	10.89	–	0.843	–
QT*	14.78	1.08	9.45	–	–	14.23	1.03	9.55	–	–	–
SummPip*	13.05	1.15	8.90	–	–	12.96	1.37	9.32	–	–	–
Copycat*	14.05	1.56	10.25	–	–	12.47	1.31	9.41	–	–	–
OPINESUM	31.58	4.58	23.79	9.03	0.843	27.88	4.42	21.45	7.81	0.834	–
CPSum	33.56	**10.82**	27.78	13.00	0.867	29.81	**9.67**	24.57	**11.47**	**0.855**	$O(\|\mathbb{D}\|^2)$
EvioSum	**34.21**	10.77	**29.91**	**13.05**	**0.868**	**30.25**	8.26	**25.99**	10.90	**0.855**	$O(\|\mathbb{D}\|)$
Supervised methods:											
BART-base	38.33	12.48	29.49	15.18	0.848	33.88	10.73	27.22	13.06	0.830	–

Table 2: Model Comparison in Terms of Opinion Coverage

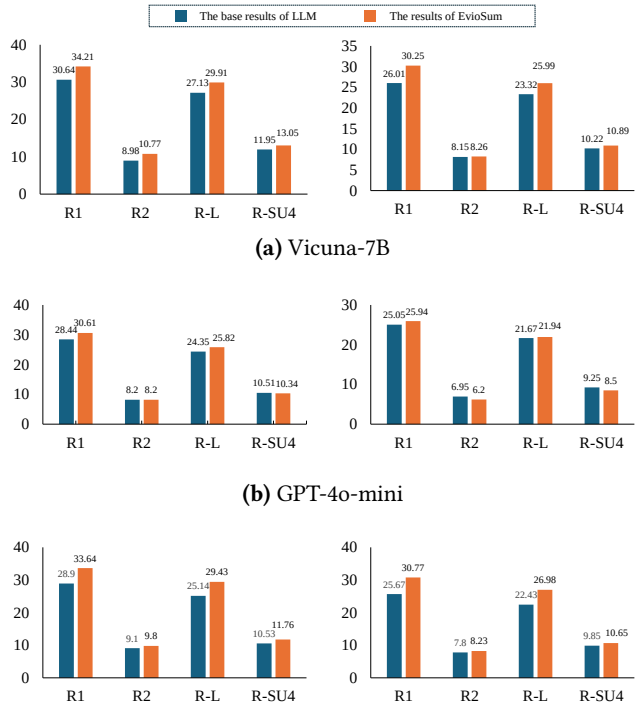
Method	Opinion Coverage against reference	Opinion Coverage against source documents
Deepseek-V3	0.791	0.858
GPT-4o-mini	0.782	0.866
OPINESUM	0.266	0.606
CPSum	0.726	0.869
EvioSum	**0.798**	**0.908**

when evaluating opinion coverage. Specifically, we separately combine the content redundancy with each of the two perspectives of opinion coverage by GEVAL [24], with prompts shown in Fig. 3.

The results in Table 2 show that our method outperforms the comparison methods on both perspectives of opinion coverage. These observations highlight the impact of the document partition and opinion extension mechanisms in the summary generation process. Specifically, the document partition organizes the interrelationships between the opinions in the source documents, and enhances the LLM's ability to focus on multiple opinions, leading to summaries with higher opinion coverage. Meanwhile, the opinion extension enhances semantic coherence and helps avoid repetitive content in the generated summaries.

5.2 The Adaptation to Different LLMs

To verify the robustness of our framework across different LLMs, we conduct experiments based on three LLMs: Vicuna-7B, GPT-4o-mini, and Claude-3.5-Sonnet. The results are presented in Fig 2, where the blue bars represent the baseline results obtained by guiding LLMs using the regular prompts in Section 4.3, and the orange bars indicate the results of our method based on the same LLM. The comparison results show that EvioSum achieves performance improvements across all LLMs, demonstrating the robustness of our framework to different backbones. Notably, compared to the



(a) Vicuna-7B

(b) GPT-4o-mini

(c) Claude-3.5-Sonnet

Figure 2: Performance of EvioSum on different LLMs

percentage of improvements on ROUGE-2 and ROUGE-SU4 scores, EvioSum achieves more substantial gains on ROUGE-1 and ROUGE-L. This suggests that EvioSum is particularly effective in improving lexical accuracy and logical coherence, while also preserving phrase-level collocations.

Table 3: Ablation Study

Model	EO Dataset				CO Dataset			
	R1	R2	RL	RSU4	R1	R2	RL	RSU4
EvioSum	34.21	10.77	29.91	13.05	30.25	8.26	25.99	10.90
w/o Majority	-0.80	-0.11	-0.72	-0.45	-0.65	-0.20	-0.90	-0.27
w/o Extension	-0.89	-0.55	-1.03	-0.37	-2.26	-0.36	-1.64	-0.55

Table 4: Comparison of EvioSum with Sel-sim and Sel-llm

Method	M	EO dataset		CO dataset	
		Win	Loss	Win	Loss
EvioSum VS Sel-sim	1	36	12	29	14
	2	42	6	38	5
	3	43	5	36	7
EvioSum VS Sel-llm	1	42	6	36	7
	2	45	3	40	3
	3	46	2	40	3

Table 5: Human Evaluation Results

Evaluation target	Comparison methods	Win	Tie	Loss
Summary	EvioSum VS CPSum	17	9	4
	EvioSum VS DeepSeek	19	7	4
Explanation Set	EvioSum VS Sel-sim	18	4	8

5.3 Ablation Study

There are two key components in our method: the majority opinion guided document partition and the opinion extension. To evaluate their contributions, we conduct two ablation experiments, with the results presented in Table 3. In the **w/o Majority setting**, we remove the guidance of majority opinion and instead directly prompt the LLM to partition the source documents into multiple evidence sets. Then the largest sets is used to generate the majority opinion by LLM, after which the remaining sets are used to extend and complement the majority opinion. In the **w/o Extension setting**, we remove the opinion extension mechanism. Instead, we generate individual opinions for each evidence set by LLM and then use the same LLM to merge these opinions into a final summary.

The results in Table 3 show that both components contribute to the model performance. Moreover, we find that the opinion extensions have a more significant effect on CO dataset than on EO dataset. One possible reason is that the COVID-19-related opinions in CO dataset tend to exhibit more clearly oppositions and stances. In contrast, the election-related EO dataset involves more nuanced, ideologically entangled, making it harder for the opinion extension component produce logically coherent summaries.

5.4 The Evaluation Results of Explanation Set

Due to the absence of comparison methods, to explore whether the *explanation set* E^* accurately reflects the underlying divergences between the opinions o_m and o_v, we construct two baseline methods, **Sel-sim** and **Sel-llm**. The **Sel-sim** method forms E^* by selecting the documents that are semantics similar to the opinions o_m and o_v, respectively, formally: $E^* = \{(E_i, E_j) \mid E_i = Top\text{-}M\{d|Sim(d, o_m)\}, E_j = Top\text{-}M\{d|Sim(d, o_v)\}$, where $Top\text{-}M\{\}$ denotes selecting the M evidence with the highest similarity. The **Sel-llm** method directly uses the GPT-3.5-turbo to select M evidence pairs from the document set $\mathbb{D}$ based on o_m and o_v.

We adopt a pairwise evaluation to compare our method with the above two methods. We present the *explanation set* generated by our method and comparison methods, along with the opinions o_m and o_v, to GPT-4 and ask it to judge which set better captures the divergence between the two opinions. The used prompt is shown in Fig. 3. We count the number of winning cases for each method. If a summary does not contain any divergent opinions, i.e., the divergent evidence set is empty, it is excluded from the results. The results in Table 4 show that our method wins over **Sel-sim** and **Sel-llm** in more cases, demonstrating its ability to faithfully capture the divergence between opinions. We can also see that the advantage of our method becomes more pronounced as the number of evidence pairs increases, since a larger M allows it to cover a wider range of aspects between o_m and o_v. In contrast, the baseline methods tend to select evidence that is semantically similar or repetitive, resulting in redundancy and limited representation of the underlying divergences.

5.5 Human Evaluation

We randomly selected 30 out of 100 samples for human evaluation and recruited three annotators. Each annotator has experience in summary evaluation. To reduce subjective bias stemming from individual differences in scoring scales, we employed a pairwise evaluation strategy. This strategy requires annotators to directly compare two system outputs and indicate which one is better, thereby facilitating more consistent and reliable judgments.

To evaluate the generated summaries, we provide each annotator with the source documents, the reference summary, and the summaries generated by two comparison methods. To guide the annotators, we clarify the evaluation requirements and design three criteria that cover multiple dimensions: the coverage of opinions in the reference, the factual consistency against the source documents, and the logical organization of the opinions. We use *Krippendorff´s alpha* coefficient to measure the inter-annotator agreement. For EvioSum VS CPSum, the value was 0.3646, and for EvioSum VS DeepSeek, it was 0.3849, both indicating fair agreement. Due to the highly subjective nature of evaluating summaries, these values already indicate an acceptable agreement[9, 32].

Table 6: Case Study

Topic: Film star Leonardo DiCaprio uses his Oscar-winning speech to raise the issue of climate change

Reference: Majority opinion: Leonardo DiCaprio is a hero for talking about climate change in his acceptance speech. We love him. Let's hope the world takes notice and takes action. Other opinions: Some think Leonardo DiCaprio is a hypocrite because he takes flights in private jets. Some think that eating animals is a bigger problem, others think over-population is a bigger problem. A minority claim climate change is a hoax. A minority believe that introducing the subject in an Oscar speech felt awkward

Generated Summaries

CPSum: There are different opinions on Leo DiCaprio's Oscar acceptance speech about climate change. Some people appreciate his use of the platform to raise awareness, while others criticize his hypocrisy for taking private jet flights. Some believe that addressing climate change is important, while others dismiss it as a convenient issue.

GPT-4o: The majority think Leonardo DiCaprio's use of his Oscar acceptance speech to address climate change was a commendable and impactful act. Many praised him for using his platform to draw attention to an important global issue rather than focusing on personal accolades. Supporters appreciated his call to action on climate change and hoped his words would inspire greater awareness and action. … A minority criticized DiCaprio's focus on climate change, with some labeling it hypocritical due to his reported use of private jets and luxury lifestyles, which they feel contradict his environmental advocacy. Others expressed skepticism, questioning the effectiveness of his message or dismissing the urgency of climate change altogether. Some found his approach awkward or poorly received...

EvioSum:

The majority opinion: The majority opinion think that Leonardo DiCaprio's speech on climate change at the Oscars was significant and impactful, as it brought attention to a pressing global issue and inspired action among viewers.

The divergent opinion: However, others question DiCaprio's commitment to climate change by pointing out his use of private airplanes, express skepticism about people actually taking climate change seriously, or criticize the integration of climate change rhetoric into speeches. Furthermore, some believe that population growth poses a similar threat to the planet as climate change, and emphasize the need for a more holistic approach to addressing environmental issues.

Explanation Set: The evidence pairs for explaining the divergence between the majority opinion o_m and the divergent opinion o_v.

Aspect 1: Climate change as a global issue

Pair 1: (I love how Leonardo DiCaprio related his speech back to climate change VS There's something awkward about Leonardo DiCaprio trying to shoehorn climate change rhetoric into his speech.)

Pair 2: (So great to see Leonardo DiCaprio talking about climate change in his acceptance speech VS There's something awkward about Leonardo DiCaprio trying to shoehorn climate change rhetoric into his speech.)

Aspect 2: Utilizing the Oscars platform for advocacy

Pair 3: (Great that Leonardo DiCaprio used the Oscars platform to address climate change VS There's something awkward about Leonardo DiCaprio trying to shoehorn climate change rhetoric into his speech.)

Green represents evidence supporting the majority opinion, while brown represents evidence supporting the divergent opinion.

The evaluation results in Table 5 indicate that EvioSum demonstrates a clear overall advantage. According to annotator feedback, compared to CPSum and DeepSeek, the advantages of EvioSum lie in its clear articulation of opinions, logical coherence, and greater conciseness. Summaries generated by DeepSeek tend to list rather than organize opinions. This illustrates a well-known limitation of LLMs: without explicit guidance, they often fail to produce well-structured outputs. CPSum leverages the feedback to calibrate the prompts, enhancing the model's attention to diverse opinions. While its conciseness is comparable to that of EvioSum, it suffers from weak logical connections between opinions in summaries.

To evaluate the *explanation set*, we provide each annotator with the majority opinion, the divergent opinion and the *explanation set* generated by two comparison methods. For each set, the annotators are asked to judge whether the divergence between the evidence accurately reflects the divergence between the majority and divergent opinions. The *Krippendorff's alpha* coefficient between the annotators is 0.704, which denotes a high inter-annotator agreement. The evaluation results presented in the Table 5 indicate that our method is able to identify more effective evidence. The Sel-sim method has two main limitations. First, the retrieved evidence is not guaranteed to support the corresponding opinion, as high semantic similarity with the opinion does not indicate stance alignment. Second, some evidence is highly relevant to both the majority and divergent opinions, which results in retrieving the same evidence for both the majority and divergent opinion.

5.6 Case Study

To visually compare different methods, we present the generated summaries in Table 6, where the co-occurring opinions in both the reference and the generated summaries are highlighted in blue.

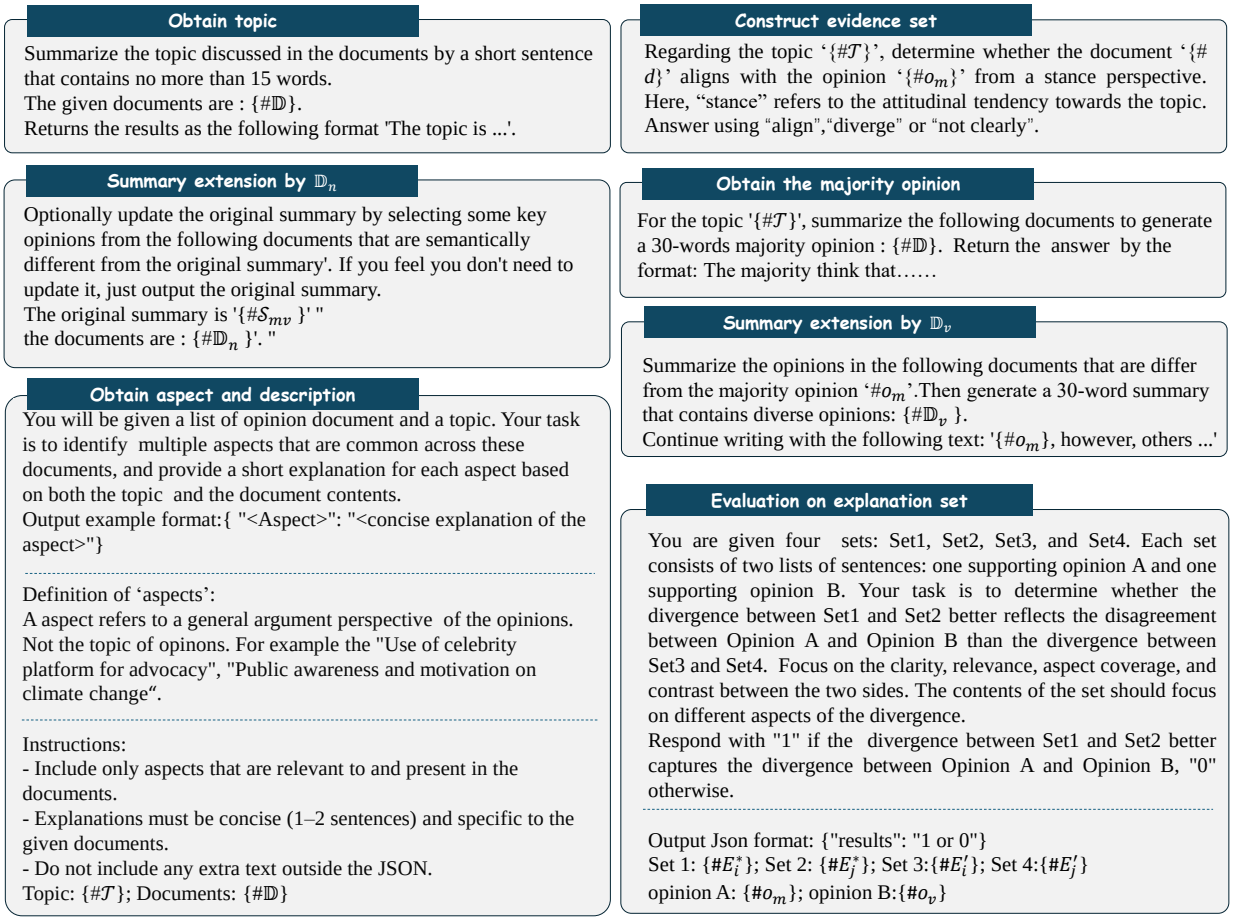


Figure 3: Prompts Used in Our Method

Compared to the baseline methods, we find that the summaries generated by EvioSum cover a greater portion of the opinions presented in the reference summaries. Some opinions such as the *'population growth'* related opinion are overlooked by other methods. However, they are successfully captured by our method. We also observe that EvioSum effectively organizes multiple opinions, and the semantic relationships among these opinions are logically coherent, which reflects the guiding roles of document partition and opinion extension in our method. We also present the *explanation set* corresponding to the summaries generated by EvioSum at the bottom of Table 6, where the high support evidence can be ranked by the function $sup()$ in Section3.3. We observe that the evidence pairs generated by our method effectively cover aspects underlying both the majority and divergent opinions. For example, it covers two important aspects of the public concern: *'limate issues'* and the *'Oscars platform'*. For each aspect, the pairs not only substantiate the respective opinions but also clearly reveal the aspect-specific representation of the divergences.

6 Conclusions

For the faithful and interpretable opinion summarization task, we propose the evidence-guided framework, which firsts generate majority opinion to partition source documents into multiple evidence sets, and then extracts opinions from each sets to form a summary. This not only improves the interpretability by revealing the association between opinions and their supporting evidence, but also ensures the semantic relevance between opinions in summaries. To explain the divergence between opinions in summaries, we select evidence from different evidence sets to construct an *explanation set* by designing a submodular optimization algorithm. Experiments on two benchmark datasets demonstrate that our summarization method outperforms multiple baselines. Both LLM and human-based evaluations demonstrate that our method identifies more comprehensive evidence that better captures opinion divergences.

7 Acknowledgments

This work was supported by the National Natural Science Foundation of China(62376138), the Innovative Development Joint Fund Key Projects of Shandong NSF(ZR2022LZH007).

8 Ethical Considerations

This work focuses on opinion summarization from publicly available text data. It does not involve collection or use of personal or sensitive information. We do not foresee negative societal impacts related to fairness, privacy, security, safety, or misuse.

References

[1] 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 (03 2021), 277–293. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00366/1924181/tacl_a_00366.pdf

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). https://api.semanticscholar.org/CorpusID:11212020

[3] 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*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada, 9282–9300.

[4] Iman Munire Bilal, Bo Wang, Adam Tsakalidis, Dong Nguyen, Rob Procter, and Maria Liakata. 2022. Template-based Abstractive Microblog Opinion Summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 1229–1248.

[5] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Online, 5151–5169.

[6] Seonglae Cho, Myungha Jang, Jinyoung Yeo, and Dongha Lee. 2024. RTSUM: Relation Triple-based Interpretable Summarization with Multi-level Salience Visualization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, Kai-Wei Chang, Annie Lee, and Nazneen Rajani (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 53–60. https://doi.org/10.18653/v1/2024.naacl-demo.5

[7] Eric Chu and Peter J. Liu. 2018. MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarization. In *International Conference on Machine Learning*.

[8] Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.* 22, 1 (dec 2004), 457–479.

[9] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409. https://doi.org/10.1162/tacl_a_00373

[10] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large Language Models are Few-Shot Summarizers: Multi-Intent Comment Generation via In-Context Learning. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering* (<conf-loc>, <city>Lisbon</city>, <country>Portugal</country>, </conf-loc>) *(ICSE '24)*. New York, NY, USA, Article 39, 13 pages.

[11] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 4952–4957. https://doi.org/10.18653/v1/D18-1537

[12] Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in Opinion Summarisation from Pre-training to Adaptation: A Case Study in Political Bias. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1041–1055. https://aclanthology.org/2024.eacl-long.63/

[13] Fariz Ikhwantri, Hiroaki Yamada, and Takenobu Tokunaga. 2024. Analyzing Interpretability of Summarization Model with Eye-gaze Information. In *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:269804831

[14] Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. ConsistSum: Unsupervised Opinion Summarization with the Consistency of Aspect, Sentiment and Semantic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. New York, NY, USA, 467–475.

[15] Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2013).

[16] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *ArXiv* abs/2205.11916 (2022). https://api.semanticscholar.org/CorpusID:249017743

[18] Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity Calibration for Opinion Summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5211–5224. https://doi.org/10.18653/v1/2024.naacl-long.291

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Online, 7871–7880.

[20] Miao Li, Jey Han Lau, and Eduard Hovy. 2024. A Sentiment Consolidation Framework for Meta-Review Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10158–10177. https://doi.org/10.18653/v1/2024.acl-long.547

[21] Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Aspect-Aware Decomposition for Opinion Summarization. arXiv:2501.17191 [cs.CL] https://arxiv.org/abs/2501.17191

[22] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[23] Yajing Liu, Edwin K. P. Chong, Ali Pezeshki, and Zhenliang Zhang. 2020. Submodular optimization problems and greedy strategies: A survey. *Discret. Event Dyn. Syst.* 30, 3 (2020), 381–412. https://doi.org/10.1007/S10626-019-00308-7

[24] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[26] Annie Louis and Joshua Maynez. 2023. OpineSum: Entailment-based self-training for abstractive opinion summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada, 10774–10790.

[27] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46534–46594. https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf

[28] Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2023. Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=ooxDOe7ZtBe

[29] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Vancouver, Canada, 1073–1083.

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *preprint* (12 2013).

[31] Haonan Wang, Gao Yang, Bai Yu, Mirella Lapata, and Heyan Huang. 2020. Exploring Explainable Selection to Control Abstractive Generation. *CoRR* abs/2004.11779 (2020). arXiv:2004.11779 https://arxiv.org/abs/2004.11779

[32] Jian Wang, Yuqing Sun, Yanjie Liang, Xin Li, and Bin Gong. 2024. Iteratively Calibrating Prompts for Unsupervised Diverse Opinion Summarization. In *European Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:273589315

[33] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh,

and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 606–615. https://doi.org/10.18653/v1/D16-1058

[34] Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Toronto, Canada, 8640–8665.

[35] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative Text Summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10644–10657. https://doi.org/10.18653/v1/2023.findings-emnlp.714

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *ArXiv* abs/1904.09675 (2019). https://api.semanticscholar.org/CorpusID:127986044

[37] Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. SummPip: Unsupervised Multi-Document Summarization with Sentence Graph Compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. New York, NY, USA, 1949–1952.

[38] Yicheng Zou, Kaitao Song, Xu Tan, Zhongkai Fu, Qi Zhang, Dongsheng Li, and Tao Gui. 2023. Towards Understanding Omission in Dialogue Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14268–14286. https://doi.org/10.18653/v1/2023.acl-long.798