Inducing Argument Facets for Faithful Opinion Summarization

Jian Wang[1], Yanjie Liang[2], Yuqing Sun*[3], Bin Gong[3]
School of Software, Shandong University
[1] wangjian026@126.com,[2]202335329@mail.sdu.edu.cn
[3]{sun_yuqing,gb}@sdu.edu.cn

Abstract

Faithful opinion summarization task refers to generating a summary for a set of documents that covers the majority and minority opinions in documents. Inspired by the cognitive science that argument facet is the focus of an opinion, we propose the facets-guided opinion summarization method (*FacSum*). By inducing the facets, we partition the documents into multiple facet-specific sets. Then key phrases are extracted as the representatives of each set and the number of facets is used for constraining the length of summary, both of which are used to guide LLMs to cover different argument facets of opinions while keeping the summary concise. We perform experiments on two representative datasets and the results show that our method outperforms the state-of-the-art (SOTA) methods and multiple LLMs. The ablation studies indicate that the introduced facets contribute to improving model performance by enabling the coverage of minority opinions while preserving the majority ones. The results based on different LLMs demonstrate that our method can improve the performance of LLMs with varying model sizes. We apply *FacSum* to the summarization of professional paper reviews, and the results confirm its effectiveness in specialty domains as well.

1 Introduction

For a topic, there are often many different opinions. Some opinions reflect the common concerns of social groups and are generally recognized as the majority opinions. In contrast, minority opinions, although expressed less frequently, are equally important since they reveal potential problems, risks, or neglected demands(Lei et al., 2024). Our faithful opinion summarization task (Li et al., 2025a; Zhang et al., 2024; Bilal et al., 2022) aims to generate concise summaries that accurately capture both the majority and minority opinions in a given opinion document set. This task is crucial in scenarios such as the public policy making to help decision-makers quickly and comprehensively understand different opinions so as to prevent decision bias.

Recently, LLMs have become the main choice for the above task (Bhaskar et al., 2023; Li et al., 2024; Zakkas et al., 2024). Some methods focus on constructing high-quality prompts for guiding LLMs (Zhang et al., 2023; Wang et al., 2024). They use a self-evaluation framework that evaluates the generated summaries by LLM and construct feedback to calibrate the prompts. While these methods can guide LLMs to generate summaries containing different opinions, they come with high computational overhead. Other methods decompose the summarization task into multiple simpler subtasks using a Chain-of-Thought (CoT) approach(Kojima et al., 2022), leveraging the strengths of LLMs on simpler tasks (Bhaskar et al., 2023; Li et al., 2025b). For example, the TCG method clusters the documents, generates a sub-summary for each cluster, and then merges all sub-summaries into a complete summary(Bhaskar et al., 2023). These methods improve the semantic coverage of summaries to source documents. However, due to the lack of explicit guidance for minority opinions, some marginal opinions are easily overlooked during the clustering and generation processes.

To generate summaries that cover both majority and minority opinions while maintaining concise, we propose the **Fac**et-guided opinion **Sum**marization method (**FacSum**). Here, a *'facet'* refers to a specific latent statement dimensions of an opinion (Li et al., 2024; Vázquez Campos and Liz Gutiérrez, 2015), such as for the topic *"Should schools offer cash bonuses for good test scores "*, the facets including *"Doubts about practicality and feasibility"*, *"Concerns about bribery and ethics"*. FacSum explicitly highlights the latent facets within the documents to construct prompts

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16153–16166
November 4-9, 2025

that can guide the attention of LLMs. The facet-sensitive prompt includes two kinds of guidance, where the semantics guidance considers the facet coverage in documents for guiding LLMs to cover both the majority and minority opinions with all facets, and the format guidance by the number of facets is used to constrain and organize opinions for reducing redundancy.

Specifically, we first use an LLM to induce the topic and facets within the document set. Then, based on the identified facet of each document, the source documents are grouped as multiple facet-specific sets. Next, key phrases are extracted to represent each subset and are subsequently merged to form a comprehensive set of facet representations. These phrases form the semantics guidance. Compared to directly extracting key phrases from the document sets, our method can capture the phrases from majority opinion while retaining those from minority opinions. The number of facets is used for constraining the maximum length of summary since the more facets involved in documents, the more content in summary. To guide LLMs in organizing summaries in a logical manner, we design the narration template, which is used together with the above guidance. We perform experiments on two social datasets and a scientific dataset. The results demonstrate that our method outperforms the SOTA methods and several LLMs.

2 Related Work

For the opinion summarization task, since the annotated data is difficult to obtain, the mainstream methods have focused on unsupervised settings and can be categorized into two types. The first type adopts the domain knowledge to pre-trains a summarization model, while the second type explores how to directly leverage the knowledge encoded in LLMs.

2.1 Pre-training Methods for Summarization

The pre-training methods incorporate domain knowledge to pre-train the model by constructing self-supervised objectives, which can be divided into two categories. The first category focuses on training model by reconstruction tasks (Chu and Liu, 2018; Suhara et al., 2020; Bražinskas et al., 2020; Angelidis et al., 2021; Basu Roy Chowdhury et al., 2022). For example, MeanSum generates a summary by decoding from the average semantic representation of multiple input documents, where the representations are learned by reconstructing the input documents(Chu and Liu, 2018). The HERCULES method represents sentences in opinions as paths through a hierarchical discrete latent space, extracts common subpaths shared across multiple opinions, and decodes them to generate the summary(Hosking et al., 2023).

The second category involves constructing pseudo-summaries from the input documents(Ke et al., 2022; Reinald Kim Amplayo, 2021; Louis and Maynez, 2023) and uses these summaries to train a model. ConsistSum obtains the aspect and sentiment distribution for each document. Then it chooses the central document as the pseudo-summary based on the distance between the distributions (Ke et al., 2022). OPINESUM method selects multiple high entailed sentences in documents using the textual entailment model. The selected sentences are concatenated as pseudo-summaries(Louis and Maynez, 2023). The PELMS method generates pseudo-summaries by selecting a set of representative sentences that cover frequently occurring topic clusters within the input documents (Peper et al., 2024).

The above two types of methods focus on consensus opinions and lack attention to the minority opinions, making them difficult to apply to our summarization task. There are also some methods aim to cover multiple opinions in documents(Li and Chaturvedi, 2024; Bar-Haim et al., 2021), such as the TokenCluster method(Li et al., 2023), which constructs the summary by selecting the sentences that can cover the the general opinion of each aspect. The KPA method extracts key points from a collection of reviews and establishes a mapping from opinions to these key points, thereby assigning a weight to each key point and selecting those with higher weights to form the summary(Bar-Haim et al., 2021). These methods encompass multiple opinions. However, they are only applicable to review scenarios with a single theme, and the generated summaries lack comprehensiveness.

2.2 LLM-based Methods for Summarization

These methods focus on effectively leverage the knowledge of LLMs and can be categorized into two classes. The first class aims to construct high-quality prompts to guide LLMs directly, whereas the second adopts a Chain-of-Thought (CoT) approach (Kojima et al., 2022), guiding the LLM through a step-by-step summarization process.

For the first class, methods such as SumCOT

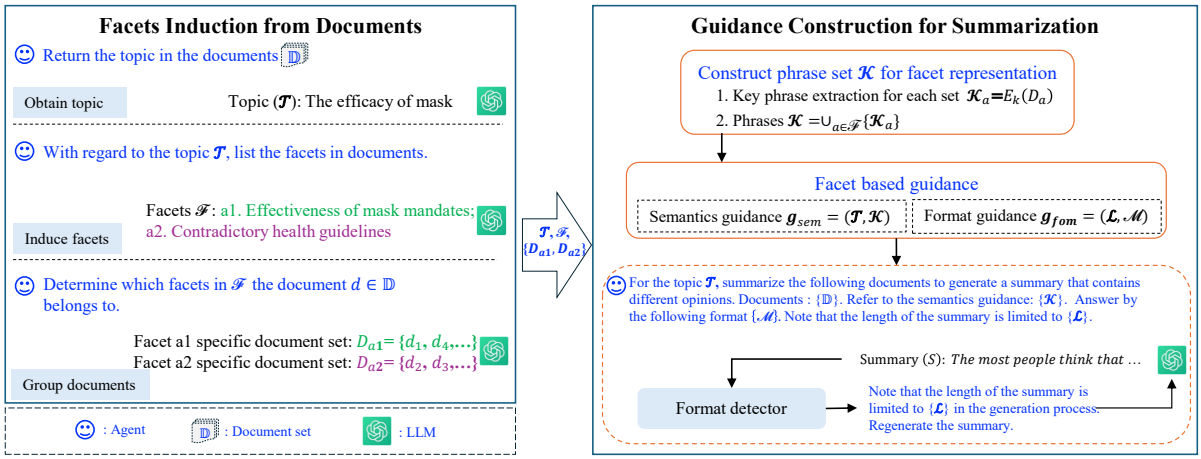


Figure 1: The framework of our facet-guided summarization method. The agent in this framework refers to an autonomous module that applies the designed prompts to guide LLMs.

(Wang et al., 2023), SummIT (Zhang et al., 2023) and CPSum (Wang et al., 2024) iteratively construct feedback by evaluating the generated summaries and update prompt for generating new summary. For instance, CPSum selects high-quality sentences in generated summaries as the guidance, adds them into the prompt for guiding LLM to retain the semantics of these sentences while exploring new opinions (Wang et al., 2024). These methods perform well due to the explicit guidance in the prompts. However, due to the large search space, they usually iterate multiple times and have a large computational overhead.

For the second class, the related methods split the summary generation process into multiple phases and construct prompts for each phase (Bhaskar et al., 2023; Li et al., 2025c, 2024). For example, the TCG method first clusters sentences based on the aspects, and each cluster is then chunked for generating multiple chunk-level summaries. Subsequently, the final summary for an aspect is generated by merging all chunk-level summaries (Bhaskar et al., 2023). Although these methods achieve high coverage of the source documents, they overlook the semantic connections between chunks, resulting in an isolated stacking of opinions and a lack of logic in the generated summaries.

3 Problem and framework

Given a document set $\mathbb{D} = \{d_1, d_2, \cdots, d_{|\mathbb{D}|}\}$, the faithful opinion summarization task aims to generate a summary $\boldsymbol{S}$ that captures both majority and minority opinions within $\mathbb{D}$. For this task, we propose the facet-guided summarization method (**FacSum**) that induces the facets in $\mathbb{D}$ for guiding LLMs to focus on these opinions. We use *'facet'* to denote the fine-grained argumentative dimensions of an opinion (e.g., Vaccination Safety for the Elderly for the topic 'Vaccination') (Li et al., 2024), as opposed to broader *aspect* (e.g., Vaccination Risks) used in prior works (Bhaskar et al., 2023; Angelidis et al., 2021), allowing for a finer delineation of opinion diversity. As shown in Fig.1, our method contains two parts: to induce the facets from documents and to construct the facet related guidance for summarization.

We first induce the facets contained in documents by LLMs, based on which the documents $\mathbb{D}$ are grouped into multiple facet-specific sets. As a result, each document set correspond to a facet. Then key phrases are extracted from each set of documents as the representative of the corresponding facet. We merge all the key phrases from each facet-specific set together for constructing the prompt that contains the semantics and format guidance. The semantic guidance encourages the LLM to attend to all facets of opinion, while the format guidance helps organize the multiple opinions in a structured manner to reduce redundancy.

3.1 Facets induction from documents

We first obtain the topic $\boldsymbol{\mathcal{T}}$ discussed in documents $\mathbb{D}$ so that we can induce the facets that focus on the same topic, formally, $\boldsymbol{\mathcal{T}} = P_{llm}(p_t, \mathbb{D})$, where $P_{llm}()$ means a LLM and p_t is a pre-designed prompt. Then the facet set $\mathcal{F}$ is induced by another prompt p_f, i.e., $\mathcal{F} = P_{llm}(p_f, \mathbb{D}, \boldsymbol{\mathcal{T}})$. The details of p_t and p_f are given in Appendix D.

For each facet $a \in \mathcal{F}$, we use LLM to check whether document $d \in \mathbb{D}$ belongs to facet a , de-

noted by $\mathbb{I}(d,a) \in \{0,1\}$. $\mathbb{I}(d,a) = 1$ indicates that the opinion in document d belongs to the facet a. Otherwise, $\mathbb{I}(d,a) = 0$. Based on the check results, all the documents in $\mathbb{D}$ are partitioned into multiple facet-specific document subset. A subset D_a corresponding to a is defined as:

$$D_a = \{d \in \mathbb{D} | \mathbb{I}(d,a) = 1\} \quad (1)$$

3.2 Guidance construction for summarization

To introduce the facet information for constructing LLM prompts, we use a key phrase extraction method denoted as $E_k()$ to extract the key phrase set $\boldsymbol{\mathcal{K}}_a$ from each D_a:

$$\boldsymbol{\mathcal{K}}_a = E_k(D_a) \quad (2)$$

We adopt *TextRank* (Mihalcea and Tarau, 2004) as $E_k()$ since it offers higher efficiency and better interpretability than the common used methods such as PromptRank(Kong et al., 2023), EnsembleKEM(Abibullayeva et al., 2024). The key phrases from each facet-specific document set are merged together as the *diversity-oriented phrase set* $\boldsymbol{\mathcal{K}}$ to represent all the facets in $\mathbb{D}$, where U is the union operation of sets:

$$\boldsymbol{\mathcal{K}} = \mathrm{U}_{a \in \mathcal{F}} \{\boldsymbol{\mathcal{K}}_a\} \quad (3)$$

Based on $\boldsymbol{\mathcal{K}}$, we form the semantics guidance $\boldsymbol{g}_{sem} = (\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{K}})$ that contains the topic $\boldsymbol{\mathcal{T}}$ and the *diversity-oriented phrase set* $\boldsymbol{\mathcal{K}}$. The interpreted topic $\boldsymbol{\mathcal{T}}$ is used to avoid semantic bias in summarization process, and the $\boldsymbol{\mathcal{K}}$ can guide LLM to focus on multiple opinions with different facets.

In general, the more facets the document set $\mathbb{D}$ contains, the longer the summary length. However, aiming for high facet coverage may easily lead to content redundancy, resulting in excessively long summaries (Alguliev et al., 2012; Srivastava et al., 2022; Xiao and Carenini, 2020). To solve the above problems, we calculate the maximum length of the summary based on the number of facets in $\mathbb{D}$, denoted as $\boldsymbol{\mathcal{L}} = |\mathcal{F}| * l$, where l is a hyper-parameter representing the maximum number of words for summarizing a single facet. To organize opinions and reflect the intention that summaries need to contain both the majority and minority opinions, we also design the template $\boldsymbol{\mathcal{M}}$ that can be automatically adapted according to different expression needs. Here, we generate multiple templates using a LLM and select one through expert evaluation. The length constrain $\boldsymbol{\mathcal{L}}$ and template $\boldsymbol{\mathcal{M}}$ form our format guidance $\boldsymbol{g}_{fom} = (\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{M}})$.

We use the task statement I to organize the two kinds of guidance and obtain the prompt $\boldsymbol{p} = (I, \boldsymbol{g}_{sem}, \boldsymbol{g}_{fom})$ for summarization. A complete prompt is shown at the bottom of the second part in Fig.1. In order to ensure that the summary conforms to the guidance, we detect the format of the generated summaries. If the summaries do not conform to the guidance, additional reminder text are added to prompt for regenerating.

4 Experiments

4.1 Datasets and metrics

We use the benchmark Microblog Opinion Summarization datasets (MOS) (Bilal et al., 2022) for experiments. MOS is collected from the Twitter site [1], which includes two datasets with different topics: the **EO** dataset contains documents related to the election topics, while the **CO** dataset focuses on COVID-19 related topics. Each sample in the two datasets includes multiple documents and a manually generated summary. Each summary explicitly includes annotations of the majority and minority opinions.

We use the *gram*-based metric ROUGE, the semantics-based metric BERTScore (Zhang et al., 2019) to evaluate the similarity between the generated summary and reference summary. For the opinion coverage evaluation, we use the LLM based metric GEVAL (Liu et al., 2023), which evaluates summaries by constructing coverage-focused prompts. GEVAL exhibits limitations in scenarios where the generated summary inverts the roles of majority and minority opinions, as it fails to adequately reflect the severity of such semantic inconsistencies. To address this issue, another metric, GroupDif (Wang et al., 2024), calculates the similarity between the majority opinion in the generated summary and that in the reference summary, as well as the similarity between their respective minority opinions. These two similarities are then combined to produce an overall similarity score. The detailed calculation process for GroupDif is provided in Appendix C.

4.2 Implementation details

The criteria for selecting our backbone model are open-source and capable of local deployment. Since Vicuna-7B (Chiang et al., 2023) has better performances across multiple tasks compared to other models with the same size of parameters, we

[1] https://twitter.com

	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
Extractive method	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	–	–
Generative method	GPT-4o	31.80	10.20	27.53	11.86	0.861	26.26	7.98	22.85	9.58	0.849
	Claude	32.10	10.35	28.21	12.17	0.863	27.68	9.00	24.62	10.57	0.852
	DeepSeek-V3	32.14	10.15	27.92	11.99	0.863	25.62	7.68	22.22	9.61	0.852
	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
	FacSum	**35.91**	**11.83**	**30.97**	**13.82**	**0.868**	**31.42**	9.24	**27.16**	**11.59**	**0.857**
Supervised method	BART-base	38.33	12.48	29.49	15.18	0.848	33.88	10.73	27.22	13.06	0.830

The results marked with '*' are taken from the paper (Bilal et al., 2022). R-1, R-2, R-L and R-SU4 denote ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-SU4 (Lin, 2004), respectively. BS stand for BERTScore (Zhang et al., 2019).

Table 1: Model comparison results.

use Vicuna-7B (Chiang et al., 2023) as the base model of FacSum and set the temperature coefficient to 0.7. To prevent introducing a large number of irrelevant phrases when constructing semantics guidance, we extract at most 2 key phrases for each facet-specific document set. The hyperparameter l represents the number of words used to summarize one facet content. The larger the value of l, the more informative but verbose. In linguistics, clearly and briefly expressing an opinion facet typically requires approximately 15 to 20 words. To ensure the summary is concise, we set l to 15. We use *'The majority think.... A minority...'* as template $\mathcal{M}$. All experiments are conducted on V100 GPUs and all the used prompts in FacSum are shown in Appendix D. We repeat the experiment 7 times and take the mean value as the final result.

4.3 Comparison methods

We compare FacSum with two unsupervised extractive methods and multiple unsupervised generative methods. The extractive methods include the *PageRank* based method **LexRank** (Erkan and Radev, 2004) and the Vector-Quantized Variational Autoencoders based method **QT** (Angelidis et al., 2021). For the generative methods, **SummPip** (Zhao et al., 2020) builds a sentence graph for clustering documents and compresses each cluster to generate the final summary. **Copycat** (Bražinskas et al., 2020) uses a hierarchical variational autoencoder to model the process of generating new opinions from multiple documents. **OPINESUM** (Louis and Maynez, 2023) adopts the textual entailment model to obtain pseudo-summaries for training summarization model. **Claude-3.5-haiku** [2], **GPT-4o** [3] and **DeepSeek-V3** [4] are LLMs with varying model sizes. **CPSum** (Wang et al., 2024) is the SOTA method on EO and CO datasets. It proposes a self-evaluation framework that adopts the reinforcement learning mechanism to calibrate prompts. **BART-base** (Lewis et al., 2020) is the supervised model trained on EO and CO datasets. **FacSum** is our facet-guided opinion summarization method.

4.4 Main results

We show the comparison results in Table 1. The first block shows the results of extractive methods, and the second block shows the results of generative methods, including LLMs. To enhance the competitiveness of the comparison methods, we incorporated the templates used in our method into their prompts, as shown in Appendix D.

We can see that FacSum outperforms the current SOTA method CPSum and multiple LLMs in several metrics, demonstrating the high similarity between the generated summaries and the reference summaries. The large improvement on R-1 and R-L also indicate the ability of FacSum to capture key words while maintaining language fluency. Compared to the supervised model BART, FacSum obtains higher values on BERTScore. This indicates the semantics and logic advantage of the summaries generated by FacSum. The fact that BART gets low values on these semantics-based metrics also suggests that the supervised method

[2]https://www.anthropic.com/claude/haiku
[3]https://openai.com/gpt-4
[4]https://www.deepseek.com/

Method	GroupDif ↑	GEVAL (GPT-4)	GEVAL (GPT-3.5)
OPINESUM (406M)	0.66	0.266	0.606
Vicuna (7B)	0.74	0.501	0.793
GPT-4o (175B)	**0.87**	0.644	0.866
CPSum (7B)	0.78	<u>0.726</u>	**0.869**
FacSum (7B)	<u>0.79</u>	**0.760**	**0.869**

Table 2: Model comparison results in terms of opinion coverage.

typically focuses on the word level and ignores the overall semantics of the summary.

We observe that all methods obtain comparable BERTScore values, which is reasonable since BERTScore emphasizes semantic similarity at the token level. As long as the generated summaries cover semantically relevant content, even with different opinion structures or focuses, they can achieve similar BERTScore results. Notably, our method achieves the highest BERTScore, suggesting that it preserves strong semantic alignment with the reference summaries.

5 Model Analysis

5.1 Summary evaluation in terms of opinion coverage

The generated summaries should contain both the majority and minority opinions in source documents. As stated in section 4.1, we use GEVAL (Liu et al., 2023) and GroupDif (Wang et al., 2024) to measure the opinion coverage of the generated summary against the source documents and the reference summary, respectively. For GEVAL, the designed prompt for evaluation is shown in Appendix D. The calculation process of GroupDif is in Appendix C.

The comparison results in Table 2 show that FacSum outperforms the SOTA method CPSum and multiple methods on GroupDif and GEVAL, which demonstrates that the generated summaries by FacSum have high opinion coverage with respect to the source documents and the reference. To be noted here, FacSum has lower time overhead than the SOTA method CPSum since CPSum needs to determine the relationship between each document and each sentence in the summary several times.

5.2 Human evaluation

In this section we conduct a comprehensive evaluation by humans. We randomly selected 30 samples from a total of 100 for human evaluation and recruited three evaluators with experience in summarization tasks. A pairwise evaluation mechanism is employed to compare FacSum with two representative methods GPT-4o and CPSum (Wang et al., 2024). We provide the topic, the reference summary and the generated summaries by two comparison methods to all evaluators. The evaluators are asked to refer to the content of the reference and select the highest quality from the two summaries.

Comparison methods	Win	Tie	Loss
FacSum VS CPSum	18	8	4
FacSum VS GPT-4o	14	9	7

Table 3: Results of human evaluation.

To guide evaluators in focusing on the core quality indicators of the summary, we provide evaluation criteria across three dimensions: the coverage of both the majority and minority opinions in the reference, the factual consistency with the reference summary, and the logical organization of the majority and minority opinions.

We use Krippendorff´s alpha coefficient to measure the inter-annotator agreement. For FacSum VS CPSum, the value was 0.592, and for FacSum VS GPT-4o, it was 0.511, both indicating moderate agreement. Considering the discrete nature of the label space, we additionally employed Fleiss's Kappa as a supplementary measure, which yielded values of 0.580 and 0.500, respectively, also reflecting moderate agreement.

The human evaluation results in Table 3 show that, compared with CPSum and GPT-4o, FacSum achieves more wins, indicating that its outputs align better with human preferences. Since evaluating a summary involves multiple complex factors, we also observed a considerable number of ties. This highlights the difficulty of opinion summarization evaluation.

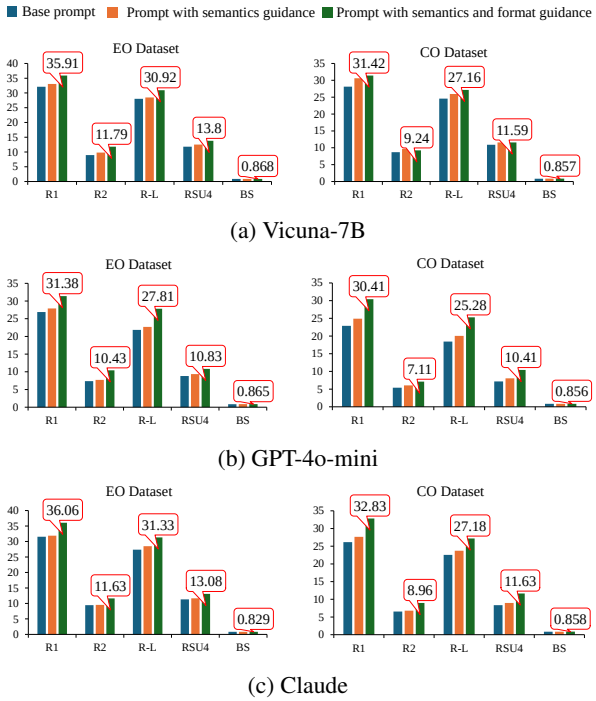
5.3 The adaptation to different LLMs

To explore the adaptation of FacSum on different LLMs, we conduct experiments based on Vicuna-7B, GPT-4o-mini and Claude[5]. For each LLM, we compare the performance of three prompts: the base prompt that only contains the task statement, the prompt with semantics guidance, and the prompt with both semantics and format guidance. We choose *"Summarize the following documents to generate a summary that contains the majority and minority opinions. Documents:* $\mathbb{D}$*"* as the base

[5]https://www.anthropic.com/news/claude-3-5-sonnet

Settings	EO Dataset				CO Dataset			
	R1	R2	RL	RSU4	R1	R2	RL	RSU4
FacSum	35.91	11.83	30.97	13.82	31.42	9.24	27.16	11.59
$w/o\ g_{sem}$	-0.93	-0.55	-1.31	-0.63	-1.72	-0.11	-1.13	-0.26
$w/o\ g_{fom}$	-2.40	-1.39	-1.79	-0.82	-1.52	-0.26	-1.16	-0.39
$w/o\ \mathcal{K}$	-0.81	-0.42	-0.71	-0.47	-1.18	-0.48	-0.76	-0.02
$w/o\ \mathcal{L}$	-0.91	-0.34	-0.80	-0.27	-1.44	-0.39	-1.06	-0.32
$w/o\ (\mathcal{L}, \mathcal{K})$	-1.13	-0.85	-1.04	-0.76	-1.45	-0.55	-1.19	-0.45
$w/o\ \mathcal{M}$	-1.26	-1.14	-0.87	-0.54	-1.68	-0.04	-1.14	-0.14

Table 4: Ablation study.



(a) Vicuna-7B

(b) GPT-4o-mini

(c) Claude

Figure 2: Results on different LLMs.

prompt. Compared to the base prompt, we can see that introducing semantics guidance improves the performance of all LLMs. Furthermore, adding format guidance resulted in additional improvements. This highlights the effectiveness of the FacSum framework and its applicability across different LLMs. It is worth emphasizing that the **Claude model based FacSum achieves new SOTA results** on both datasets. Considering the computational cost, we continue to use Vicuna as the backbone model for subsequent analyses and comparisons.

5.4 Ablation study

We first explore the effects of the two types of guidance on the model performance. The results are shown in Table 4, where $w/o\ g_{sem}$ and $w/o\ g_{fom}$ indicate that we remove the semantics guidance and format guidance, respectively. Both types of guidance play an important role, with format guidance being more significant on the EO dataset. We also examine the effects of the *diversity-oriented phrase set* and the length constraints, denoted as $w/o\ \mathcal{K}$ and $w/o\ \mathcal{L}$ in Table 4, respectively. We can see that removing the two components results in performance drop, highlighting the importance of introducing facets for both semantic and format guidance. To show the effect of template $\mathcal{M}$, i.e., $w/o\ \mathcal{M}$, we remove the template in prompts. The results show that the template $\mathcal{M}$ contributes to the performance as it explicit reflects the summarization objective. Notably, even without the template, FacSum still outperforms several baselines. A more detailed analysis of templates is given in the following section.

The main contribution in FacSum is the way of constructing the semantic guidance for the LLM prompts, which uses the key phrases extracted from the original documents. We explored two variants to replace the method of constructing semantic guidance. One variant directly uses the extracted facets as semantic guidance, referred to as **FacSum-Fac**. Another variant first partitions the source documents into multiple facet-specific subsets to generate facet-level summaries, and then extracts key phrases from these summaries as semantic guidance, denoted as **FacSum-FacKP**. The comparison results are shown in Table 5. We can see that FacSum outperforms its variants. The reason for the poor performance of FacSum-Fac is the facets being more abstract than the key phrases used in FacSum, while the reason for FacSum- FacKP is that its key phrases miss some important details.

5.4.1 Analysis on different templates

To verify that the performance gains from the template stem from its semantic guidance rather than the specific words, we replaced the keywords in

Method	EO Dataset				CO Dataset			
	R1	R2	RL	RSU4	R1	R2	RL	RSU4
FacSum	35.91	11.83	30.97	13.82	31.42	9.24	27.16	11.59
FacSum-Fac	33.33	9.54	28.69	11.98	29.18	7.94	25.47	10.38
FacSum-FacKP	31.84	8.89	27.73	11.14	29.72	7.74	26.40	10.48

Table 5: Comparison results of different variants.

Template	EO dataset				CO dataset			
	R1	R2	RL	RSU4	R1	R2	RL	RSU4
The majority think..., A minority...	35.91	11.83	30.97	13.82	31.42	9.24	27.16	11.59
The most people think..., others.....	35.98	10.62	30.88	13.43	31.42	9.0	27.15	11.55

Table 6: The effects of templates on model performance.

Phrase set	EO dataset			CO dataset		
	P	R	F1	P	R	F1
$\mathcal{K}$	**0.70**	**0.38**	**0.46**	**0.31**	**0.21**	**0.23**
$\mathcal{K}^*$	0.51	0.25	0.32	0.28	0.17	0.19

Table 7: The comparison results of different phrase sets.

the template with semantically similar alternatives. Specifically, we replace the template *"The majority..., a minority think..."* with the template *"The most people..., others "*. The results in Table 6 indicate that the substitution of keywords did not lead to a significant change in performance on the two datasets. This finding demonstrates the model's robustness to surface-level variations in the template and further supports the conclusion that the observed improvements are primarily driven by the semantics of templates rather than some specific words.

5.4.2 Evaluation of the semantics guidance

The *diversity-oriented phrase set* $\mathcal{K}$ is a key component of guidance in our method. To assess the quality of the set $\mathcal{K}$, we design the following metrics: let $\mathcal{C}$ denote the extracted phrase set in reference summary, we calculate the precision P, recall R and $F1$ for the $\mathcal{K}$, defined as $P = |\mathcal{K} \cap \mathcal{C}|/|\mathcal{K}|$, $R = |\mathcal{K} \cap \mathcal{C}|/|\mathcal{C}|$ and $F1 = (2 * P * R)/(P + R)$. We choose anther phrase set $\mathcal{K}^*$ for comparison, where $\mathcal{K}^*$ is directly extracted from the source documents using the keyword extraction method *TextRank*. The comparison results in Table 7 show that the *diversity-oriented phrase set* $\mathcal{K}$ outperforms $\mathcal{K}^*$ on all three metrics, highlighting its effectiveness in guiding the generation of high-quality summaries.

6 Adapting FacSum to Specialty Domains

We explore whether FacSum can be applied to summarization tasks in specialized domains, such as the summarization task on paper reviews. Paper reviews typically employ formal, academic language characterized by logical reasoning and analytical rigor, which differs from the social documents in EO and CO datasets. In this setting, our task aims to summarize multiple paper reviews into a meta-review that contains the consensuses and controversies (Zeng et al., 2023). We perform experiments on the ORSUM dataset (Zeng et al., 2023), which is constructed by collecting open-sourced papers and meta-reviews from Open-Review[6]. The dataset contains 15062 meta-reviews and 57536 reviews from 47 conference venues.

6.1 Implementation details for the paper review summarization task

Compared to the facets in social datasets, those in paper peer reviews are more fixed and well-defined. Therefore, we directly construct the facet set, which includes: *'Originality and Contribution', 'Clarity of Research Question', 'Related Work', 'Methodology', 'Experimental Design', 'Data Handling', 'Structure and Clarity of Writing', and 'Reproducibility and Transparency'*. Since the facets are predetermined, we compute the facet-adaptive maximum length based on the number of non-empty facet-specific document sets. To control computational overhead, we choose the first 100 samples in the ORSUM dataset for our experiments. The model's parameter settings remain consistent with those in the social domain experiments.

[6]https://openreview.net/

6.2 Comparison results on ORSUM dataset

To verify the effectiveness of our method FacSum, we compare it with the methods described in Section 4.3. The experimental results are shown in Table 8, where IC (Invocation Count) indicates the complexity measured by the number of times the LLM is invoked. The comparison results show that, compared to LLMs, FacSum achieves the best performance on all evaluation metrics. We also present a case study in Appendix B. By comparing the facets involved in *diversity-oriented phrase set* with the reference meta-reviews, we observe that the extracted phrase set covers all facets mentioned in the meta-reviews, demonstrating its effectiveness in guiding LLMs. Moreover, FacSum outperforms CPSum on several metrics while incurring lower LLM invocation complexity, highlighting its value in real-world applications.

Settings	ORSUM Dataset					
	R1	R2	RL	RSU4	BS	IC
LexRank	31.44	6.60	27.34	11.13	0.846	–
Claude	32.13	7.29	28.68	11.68	0.851	O(1)
DeepSeek-V3	27.82	5.66	25.29	9.74	0.839	O(1)
GPT-4o	29.76	6.03	26.48	10.63	0.849	O(1)
CPSum	33.97	**8.32**	30.35	**13.03**	0.858	O($\|\mathbb{D}\|^2$)
FacSum	**34.17**	8.30	**30.63**	12.85	**0.861**	O($\|\mathbb{D}\|$)

Table 8: Comparison results on paper review summarization task.

7 Conclusion

In this paper, to generate faithful opinion summaries, we propose the facet-guided summarization method, **FacSum**. We induce facets to construct fine-grained prompts that contain both semantic guidance and format guidance. The semantic guidance is constructed by extracting facet-specific key phrases, which guide LLMs to focus on both majority and minority opinions. Meanwhile, the number of facets is used to adjust the summary length. Together with a task-related template, this forms the format guidance to organize opinions and reduce redundancy. To validate the effectiveness of our method, we evaluate the generated summaries using multiple metrics covering different dimensions. Experimental results on two social domain datasets demonstrate that our method achieves state-of-the-art (SOTA) performance. **FacSum** can also improve the performance of multiple LLMs with varying model sizes, and the **FacSum** based on the Claude model even achieves new SOTA results. Furthermore, we apply our method to the summarization task on paper reviews, and the results indicate that **FacSum** generalizes well to the professional domain.

8 Limitations

We propose the facet-guided opinion summarization method, which follows the cognitive way on expressing and understanding opinions. By explicitly introducing facets, LLM is encouraged to focus on multiple minority opinions. We can also achieve high control over the generation process by emphasizing certain facets. While the method is designed to promote broad facet coverage and reflect diverse opinions, in practice, some facets may exhibit semantic overlap, such as the facets *"public trust in healthcare"* and *"confidence in vaccine safety"*. This can inadvertently lead to redundancy or skew the model's focus toward overlapping content. These observations highlight the necessity of a more refined facet filtering mechanism that enhances the salience of distinct facets while preserving the overall comprehensiveness.

Acknowledgments

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

References

Aiman Abibullayeva, Hüma Kiliç Ünlü, and Aydin Çetin. 2024. An ensemble keyword extraction model for news texts with statistical and graphical features. *Int. J. Softw. Eng. Knowl. Eng.*, 34(7):1047–1061.

Rasim M Alguliev, Ramiz M Aliguliyev, and Makrufa S Hajirahimova. 2012. Gendocsum+ mclr: Generic document summarization based on maximum coverage and less redundancy. *Expert Systems with Applications*, 39(16):12460–12473.

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.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key point analysis of business reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3376–3386. Association for Computational Linguistics.

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

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

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*, 10:1229–1248.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eric Chu and Peter J. Liu. 2018. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*.

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

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.

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*, WSDM '22, page 467–475, New York, NY, USA.

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*, volume 35, pages 22199–22213. Curran Associates, Inc.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoyan Bai. 2023. PromptRank: Unsupervised keyphrase extraction using prompt. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9788–9801, Toronto, Canada. Association for Computational Linguistics.

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)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

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*, pages 7871–7880, Online.

Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023. Aspect-aware unsupervised extractive opinion summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12662–12678, Toronto, Canada. Association for Computational Linguistics.

Haoyuan Li and Snigdha Chaturvedi. 2024. Rationale-based 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)*, pages 8274–8292, Mexico City, Mexico. Association for Computational Linguistics.

Haoyuan Li, Yusen Zhang, Rui Zhang, and Snigdha Chaturvedi. 2025a. Coverage-based fairness in multi-document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9801–9819, Albuquerque, New Mexico. Association for Computational Linguistics.

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)*, pages 10158–10177, Bangkok, Thailand. Association for Computational Linguistics.

Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025b. Aspect-aware decomposition for opinion summarization. *Preprint*, arXiv:2501.17191.

Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025c. Aspect-aware decomposition for opinion summarization. *Preprint*, arXiv:2501.17191.

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, 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*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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*, pages 10774–10790, Toronto, Canada.

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

Joseph Peper, Wenzhao Qiu, and Lu Wang. 2024. PELMS: Pre-training for effective low-shot multi-document 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)*, pages 7652–7674, Mexico City, Mexico. Association for Computational Linguistics.

Mirella Lapata Reinald Kim Amplayo, Stefanos Angelidis. 2021. Aspect-controllable opinion summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Atul Kumar Srivastava, Dhiraj Pandey, and Alok Agarwal. 2022. Redundancy and coverage aware enriched dragonfly-fl single document summarization. *Language Resources and Evaluation*, 56(4):1195–1227.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Margarita Vázquez Campos and Antonio Manuel Liz Gutiérrez. 2015. *The Notion of Point of View*, pages 1–57. Springer International Publishing, Cham.

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*.

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)*, pages 8640–8665, Toronto, Canada.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

Pavlos Zakkas, Suzan Verberne, and Jakub Zavrel. 2024. Sumblogger: Abstractive summarization of large collections of scientific articles. In *Advances in Information Retrieval*, pages 371–386, Cham. Springer Nature Switzerland.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024. Fair abstractive summarization of diverse perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

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*, SIGIR '20, page 1949–1952, New York, NY, USA.

A Case studies on social opinion summarization task

To visualize the quality of the summaries generated by FacSum, we compare FacSum with the SOTA method CPSum and GPT-4o by a case in EO dataset. The comparison results are shown in Table 9, where majority opinions are highlighted in red and minority opinions in blue. Compared to CPSum and GPT-4o, the summaries generated by FacSum cover a greater number of opinions

present in the reference summary. Moreover, FacSum's summaries are well-organized and free from content redundancy. In contrast, summaries produced by GPT-4o and CPSum exhibit issues such as missing opinions and duplicated content. For example, in the summary generated by GPT-4o, the majority opinion is repeated multiple times.

B Case study on paper review summarization task

We present a case study to analyze the performance of FacSum on the paper review summarization task. Figure 3 displays two summaries: one generated using the base prompt, and the other using prompts augmented with facet-guided guidance. The base prompt is *"summarize the following reviews to generate a summary that contains the consensuses and controversies."*.

By comparing the facets involved in the *diversity-oriented phrase set* with those in the meta-review, we observe that the extracted phrase set successfully covers all the facets present in the meta-review, indicating its high quality. Compared to the summary generated using the base prompt, the summary guided by our method is more semantically aligned with the meta-review, and demonstrates improvements in terms of facet coverage, length control, and structural organization.

Interestingly, we also find that the *diversity-oriented phrase set* includes some facets not reflected in the meta-review. For example, the meta-review does not cover the facet *"Structure and Clarity of Writing"*, while the facet *"writing"* appears in the *diversity-oriented phrase set*. This suggests that certain facets, while present in the documents, are not considered essential for summarization and should be filtered out. These observations highlight the need for a more sophisticated facet filtering mechanism to further enhance the precision and relevance of the guidance.

C The details of the evaluation metric GroupDif

GroupDif separately evaluates the similarity between the majority opinions in the reference and generated summaries, as well as the similarity between their minority opinions. It then combines these two similarity scores to produce the final evaluation result (Wang et al., 2024). The calculation process is as follows:

GroupDif splits the generated summary $\boldsymbol{S}$ into multiple opinions according to the conjunctions, transitions, and periods. For each opinion x in $\boldsymbol{S}$, GroupDif calculates the tendency value by:

$$w(x) = I(x) * max_{m,n \in R}\{sim(x,m), sim(x,n)\} \tag{4}$$

where m and n are the majority and minority opinions in reference summary R, and $sim()$ is the similarity function, $I(x)$ denotes the direction of the tendency, i.e., if $sim(x,m) > sim(x,n)$, $I(x) = 1$, otherwise, $I(x) = -1$.

Based on the direction of tendency, all opinions in the generated summaries are categorized into two groups A, B, where the opinions in group A tend to the majority opinion and the opinions in group B tend to the minority opinions. Define the tendency of each group as the mean of the tendency values of the opinions it contains:

$$score(A) = mean_{x \in A}\{w(x)\} \tag{5}$$

For a high-quality summary, $score(A)$ should converge to 1 and $score(B)$ should converge to -1. GroupDif combines these two expectations and evaluates the summary as the difference between them:

$$GroupDif(\boldsymbol{S}) = score(A) - score(B) \tag{6}$$

D Prompts for different purposes

We show the content of prompts used in our method in Fig. 4.

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. **Minority 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
Method	**Generated Summaries**
FacSum	The majority think that Leonardo DiCaprio's speech on climate change at the Oscars was a positive step towards raising awareness about the issue. They appreciate that he used the platform to call for action and bring the topic into the forefront. Some also liked how he related his speech back to climate change. A minority, however, think that DiCaprio's speech was awkward and that he was trying to shoehorn climate change rhetoric into his speech. Others thought that he was being hypocritical given his own carbon footprint. Some also felt that he should have talked about other important environmental issues as well.
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. His speech resonated with many as raising awareness about the urgency of the problem, especially in relation to his experiences during the production of "The Revenant." 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 theurgency of climate change altogether. Some found his approach awkward or poorly received. While many see his speech as a call to action, detractors believe there is more to be done to address the complexities of the issue beyond celebrity endorsements.

Table 9: Case study for social opinion summarization.

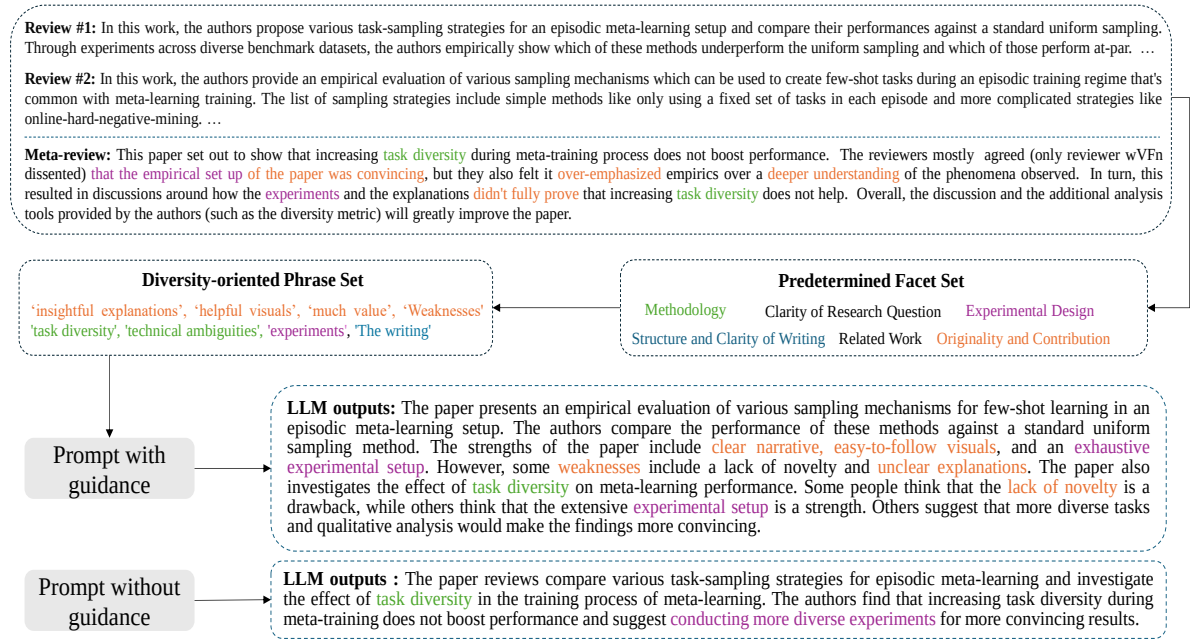


Figure 3: Case study for the peer review summarization. We use color coding to illustrate the correspondence between summary content and the predefined facets.

Prompt p_t for obtaining the topic of documents

Summarize the topic discussed in the documents by a short sentence that contains no more than 15 words.
The given documents are : {$\mathbb{D}$}.
Returns the results as the following format 'The topic is ...'.

Prompt p_f for inducing the facets in documents

With regard to the topic $\boldsymbol{\mathcal{T}}$, list the facets in the following documents. The facet refers to a specific statement dimensions of opinion.
The given documents are: {$\mathbb{D}$}.
Return results by list, the facets include:

Prompt for guiding DeepSeek, GPT-4 and Claude for summarization

Summarize the following documents to generate a summary that contains the majority and minority opinions.
The given documents are: {$\mathbb{D}$}.
Answer by the following format : 'The majority think …. A minority … ', If there is no minority opinion, only output the majority opinions.

Prompt for summary evaluation by GEVAL on EO and CO datasets

You will be given one opinion summary generated from some documents. Your task is to rate the summary on the following criteria.
Evaluation Criteria: The summary needs to contain multiple opinions with differentiation, which can be in the form of different opinions on the same topic or different perspectives on the same opinion.
Evaluation Steps:
1. Read the summary carefully and identify the key opinions.
2. Read the summary and compare it to the source documents. Check if the summary covers the majority and minority opinions.
3. Assign a score for the quality of the summary on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Source Documents: {$\mathbb{D}$}.
Summary: S.
Evaluation Form (scores ONLY): - Quality of summary:

Figure 4: Prompts used in our method.