Scientific Opinion Summarization: Paper Meta-review Generation Dataset, Methods, and Evaluation

Anonymous ACL submission

Abstract

Opinions in the scientific domain can be divergent, leading to controversy or consensus among reviewers. However, current opinion summarization datasets mostly focus on product review domains, which do not account for this variability under the assumption that the input opinions are non-controversial. To address this gap, we propose the task of scientific opinion summarization, where research paper reviews are synthesized into meta-reviews. To facilitate this task, we introduce a new ORSUM dataset covering 10,989 paper meta-reviews and 40,903 paper reviews from 39 conferences. Furthermore, we propose the Checklist-guided Iterative Introspection (CGI^2) approach, which breaks down the task into several stages and iteratively refines the summary under the guidance of questions from a checklist. We conclude through the experiments and analysis that (1) human-written summaries do not always accommodate all necessary criteria, and (2) the combination of task decomposition and iterative self-refinement shows promising discussion involvement ability and can be applied to other complex text generation using black-box LLMs.

1 Introduction

Scientific Opinion Summarization provides a succinct synopsis for scientific documents and helps readers recap salient information and understand the professional discussion. Current work on Opinion Summarization is mostly for product reviews (Hu and Liu, 2006; Amplayo et al., 2021b; Angelidis and Lapata, 2018; Suhara et al., 2020) and aims at identifying representative and consensus opinions on each aspect of interest under the assumption that the input opinions are non-controversial. However, summarizing scientific opinions is more controversial and complicated. Scientists voice agreement or disagreement for specific reasons, whereas majority voting does not always accompany consensus. Scientific meta-review summarizes the *controversies* and *consensuses* in the reviews and makes decisions.

Domain	Reviews	Meta-reviews
Product	I love these protein bars in the vanilla flavor. They taste like Rice Krispies treats with vanilla frosting ... || Nugo bars are great for breakfast, lunch or a snack ... Eat them with a tall glass of water and they will keep you satisfied for hours. || ...	These bars are fantastic and taste great like a Rice Krispy treat. Good for morning, lunch or afternoon snack and a good way to get your protein in-take. They keep you full for a long time especially if you are out and about ...
Paper	It is unclear why this work is needed. Why not use ... || The paper is well written and the math seems to be sound ... The empirical evaluation of the method is not overwhelming ... || The work appears to be sound ...	Two of the reviews suggest that the technical aspects of the paper are sound, while one reviewer questions the need for the proposed approach ... While some reviewers raised concerns about ... the majority of reviewers acknowledge the ... In light of these findings, I recommend rejection ...

Figure 1: Product meta-reviews and paper meta-review have different compositions: A product meta-review presents the most prominent opinion instead of summarizing opinions, while a paper meta-review summarizes different opinions and makes recommendations.

Furthermore, most opinion summarization datasets in the product review domain for abstractive summarization systems are synthetic, redundant cut-and-paste extracts built by combining extracted snippets, or sampling a review from the collection and pretending it to be a gold-standard meta-review (Amplayo et al., 2021b). Meanwhile, opinion summarization in scientific domains remains less explored.

To address this gap, we introduce a new task of **Scientific Opinion Summarization**, where the output meta-reviews discuss the opinions in the input reviews and accordingly make decisions. Taking research paper meta-review generation as a typical scenario, we build the **ORSUM** dataset by collecting open-sourced paper reviews and meta-reviews from OpenReview [1], covering 10,989 meta-reviews

[1]https://openreview.net/

and 40,903 reviews from 39 conference venues. Compared to the synthetic datasets from product review domains, ORSUM is built upon large-scale real-world data, enabling the applications of supervised abstractive summarization methods and more fine-grained textual analysis. In addition to meta-review generation, the structured content of ORSUM, including ratings on different aspects and multi-turn discussions, will benefit a wide range of related tasks, such as review generation (Wang et al., 2020), recommendation prediction (Deng et al., 2020; Friedl et al., 2021), review rating prediction (Li et al., 2017; Chan et al., 2020), and argument pair extraction (Cheng et al., 2020).

The task of Scientific Opinion Summarization presents a distinct set of challenges, including decision consistency, comprehensive discussion involvement, and extensive evaluation requirements. (1) *Consistency in decision guidance*: Meta-review aligns with a decision, which guides the opinion selection and discussion in the meta-review. The generated scientific meta-reviews should be able to reflect these decisions. (2) *Comprehensiveness in opinion discussion*: Unlike product meta-reviews that rely on majority voting, scientific meta-reviews access both the pros and cons, as well as opinion agreement and disagreement, to evaluate the paper from the perspective of a more senior reviewer. (3) *Extensiveness in evaluation*: The assessment of a successful meta-review should explore discussion involvement, opinion soundness, and decision consistency.

To tackle the first and second challenges, we propose a Checklist-guided Iterative Introspection (CGI^2) method. CGI^2 first breaks the task into multiple steps while constantly requesting evidence to mitigate LLM's inability to follow complicated text generation instructions and their tendency to produce hallucinations. To further enhance discussion engagement, CGI^2 iteratively revises the generated meta-review based on its own feedback derived from questions in a predefined checklist. For the third challenge, we first identify the key aspects to evaluate generated meta-reviews and propose supplementary measures for this task that can be assessed using reference-free LLM-based metrics.

Our contributions include the following:

- We introduce the task of scientific opinion summarization and construct the ORSUM dataset, which contains 10,989 meta-reviews and 40,903 reviews from 39 conferences on OpenReview. It is currently the largest paper meta-review dataset.
- We propose a Checklist-guided Iterative Introspection (CGI^2) approach, which breaks down the task into several stages and iteratively refines the summary under the guidance of questions from a checklist.
- We construct a comprehensive evaluation framework for meta-review generation and assess the generation abilities of methods in different paradigms on ORSUM.

2 Related Work

2.1 Opinion Summarization

The task of opinion summarization is typically decomposed into aspect extraction, polarity identification, and summary generation (Hu and Liu, 2006). The lack of parallel data in review summaries limits the scope of methodology into the few-shot abstractive setting (Brazinskas et al., 2020a, 2022) or unsupervised extractive setting (Angelidis and Lapata, 2018; Angelidis et al., 2020; Chowdhury et al., 2022), where the aspects and sentiments from the input reviews are collected, selected, and rearranged into the output meta-reviews.

Only a few previous opinion summarization datasets (Wang and Ling, 2016) contain gold-standard summaries and can support supervised training of abstractive models (Amplayo and Lapata, 2019). Pretrained aspect-based sentiment analysis (Suhara et al., 2020), Variational Autoencoder (Brazinskas et al., 2020b; Chu and Liu, 2019; Iso et al., 2021; Isonuma et al., 2021), and Large Language Models (Bhaskar et al., 2022) enable unsupervised abstractive approaches, where the generated summaries are validated to be more fluent, informative, coherent, and concise.

To support the training and evaluation of supervised methods, recent work constructs synthetic datasets by random sampling (Shen et al., 2023), adding noise to the sampled summary to generate documents (Amplayo and Lapata, 2020), searching for relevant reviews to act as the input document set (Elsahar et al., 2021), or sampling with trained models (Amplayo et al., 2021a,b). However, some synthetic pseudo-summaries in the product review area are known to be detached from real-world distributions, possibly irrelevant or inconsistent with input documents, and ignore salient latent aspects.

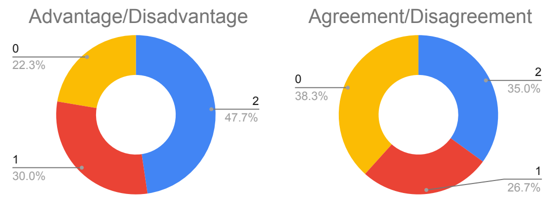

Figure 2: Meta-review composition. The scores range from 0 to 2: 0 indicates that the meta-review does not address the discussion at all. 1 signifies that the meta-review incorporates the discussion but lacks concrete evidence. 2 denotes that the meta-review involves a detailed discussion. Only 47.7% and 35.0% of meta-reviews meet the fundamental criteria for discussions of advantages and disadvantages, and consensus and controversy, respectively.

2.2 Meta-review Generation

The first attempt to generate paper meta-reviews is MetaGen (Bhatia et al., 2020), which generates an extractive draft and then uses a fine-tuned model for decision prediction and abstractive review generation. Kumar et al. (2021) emphasize decision awareness and propose a model for decision prediction and subsequent meta-review generation. The most similar work to ours is MReD (Shen et al., 2022), where 7,089 paper meta-reviews from ICLR 2018 - 2021 are manually annotated with their sentence-level structure labels. The difference is that they focus on structure-controlled text generation while our work enables scientific opinion summarization with a larger corpus, a prompting-based solution, and broader evaluations. Note that while there are other concurrent efforts to collect paper meta-reviews or reviews (Dycke et al., 2023), we are the first to model meta-review generation as scientific opinion summarization and offer a unified dataset covering more conference venues.

3 Task Formulation

Given the title, abstract, and a set of reviews from distinct reviewers of one research paper, the goal of **Scientific Opinion Summarization** is to generate a meta-review summarizing the opinions in the independent reviews and make a recommendation decision.

As noted by the area chair guidance[2], meta-review summarizes reviews by aggregating opinions to support the decision. It entails summarizing the key strengths and weaknesses of a paper, and explicitly evaluating whether the strengths surpass the weaknesses or the reverse. The meta-review also aggregates the final opinions of the reviewers after comprehensive discussions and offers an overall evaluation.

[2] https://aclrollingreview.org/aetutorial

4 ORSum Dataset

4.1 Dataset Collection and Preprocessing

We collect the ORSUM dataset for scientific opinion reviews with human-written meta-reviews from OpenReview. For each paper, we collect its URL, title, abstract, decision, meta-review from the area chair, and reviews from individual reviewers. We crawl 10,989 paper meta-reviews and 40,903 individual reviews from 39 conference venues. We only keep papers with meta-reviews longer than 20 tokens and exclude comments made by non-official reviewers. Considering the diverse format and naming of related data properties across venues, we unify the format to facilitate convenient access for future research purposes. We split the dataset into train/validation/test sets with 9,890/549/550 samples, respectively.

4.2 Dataset Comparison

We empirically compare ORSUM with existing opinion summarization datasets (or their subsets) with gold-standard summaries, including The Rotten Tomatoes (RT) (Wang and Ling, 2016), Copycat (Brazinskas et al., 2020b), OPOSUM (Angelidis and Lapata, 2018), Yelp (Chu and Liu, 2019), DENOISESUM (Amplayo and Lapata, 2020), PLANSUM (Amplayo et al., 2021b), and SPACE (Angelidis et al., 2021). A detailed introduction to these datasets is in the appendix.

Abstractiveness. The percentage of novel n-grams in the meta-review counts the ratio of n-grams that do not appear in the source reviews, which intuitively measures the abstractness of the summaries (Chen et al., 2021). Table 1 indicates a greater degree of content synthesis in ORSUM.

Redundancy. To examine the presence of insightful information in the input reviews, we assess redundancy using the Normalized Inverse of Diversity (NID) score (Xiao and Carenini, 2020) This score is calculated as the inverse of the diversity metric with length normalization: $NID = 1 - \frac{entropy(D)}{log(|D|)}$. A higher NID signifies greater redundancy. Table 1 shows lower redundancy in ORSUM, which can be attributed to the fact that many reviews address distinct aspects of the paper.

Dataset	Collection	Count(SRC)	Count(TRG)	Len(SRC)	Len(TRG)	Novel 4-gram	NID
RT	Human	246,164	3,731	20.57	21.4	97.10	0.1615
Copycat	AMT	480	180	42.63	54.33	89.62	0.2506
OPOSUM	AMT	600	60	43.51	67.77	85.92	0.1260
Yelp	AMT	3,200	200	65.25	61.15	93.26	0.1661
DENOISESUM	Synthetic	73282	837	24.32	26.45	94.12	0.2270
PLANSUM	Synthetic	249,844	869	42.81	97.2	91.40	0.2395
SPACE	Human	5000	1050	34.27	54.38	90.38	0.1671
ORSUM	Human	40,903	10,989	376.36	141.76	99.89	0.1572

Table 1: We compare ORSUM with existing opinion summarization datasets that contain gold-standard summaries. SRC refers to the source or input reviews. TRG refers to the target or output meta-reviews. A higher novel 4-gram score suggests better abstractiveness, while a lower NID score implies less redundancy.

4.3 Composition Analysis

To investigate whether the human-authored meta-reviews in ORSUM have involved the pros and cons discussion, and opinion consensus and controversy discussion, we conduct a human annotation for meta-review composition. Three annotators are asked to access the anonymous summaries in terms of discussion engagement in advantages/disadvantages and in agreements/disagreements with the scores ranging from 0 (no involvement) to 2 (detailed involvement). Annotation instructions are shown in the Appendix.

The annotation results in Figure 2 reveal that only 20.7% of meta-reviews encompass both detailed discussions, regardless of their length. For each category, 47.7%, and 35.0% of meta-reviews meet the fundamental criteria for discussions of advantages and disadvantages, and consensus and controversy, respectively. Based on these results, we conclude that *the quality of human-written meta-reviews do not always accommodate all necessary criteria and may not be suitable for developing summarization models*.

5 Checklist-guided Iterative Introspection Method for Meta-review Generation

Motivated by the unreliability of human-written meta-reviews, we turn to applying Large Language Models (LLMs) like ChatGPT (OpenAI, 2021) despite their inability to follow complicated text generation instructions and their tendency to produce hallucinations. To this end, we propose to break the task into multiple steps while consistently requesting evidence. To enhance discussion engagement and evidence-based coherence in the meta-review generation, we further introduce a checklist-guided self-feedback mechanism. The process of Self-refinement (Madaan et al., 2023) involves the LLM iteratively revising the generated meta-review based on its own feedback. Different from prior work, our checklist-guided self-feedback mechanism uses self-feedback derived from questions in a predefined checklist.

Figure 3 illustrates our proposed Checklist-guided Iterative Introspection (CGI^2) method.

Initial Run. Given the title, abstract, and a set of reviews from distinct reviewers of one research paper, CGI^2 generates a draft of the meta-review in four steps: (1) For each individual review, we prompt the LLM to extract and rank opinions and to include sentiment, aspect, and evidence. (2) Based on the extracted opinions, we prompt the LLM to list the most important advantages and disadvantages of the paper and to list corresponding reviewers and evidence. (3) The LLM is prompted to list the consensuses and controversies in the above opinions and to include the corresponding reviewers and evidence. (4) Given the decision of acceptance or rejection, the LLM is requested to write a meta-review based on the above discussion.

Iterative Runs. With the meta-review draft from the initial four-step run, CGI^2 iteratively poses questions, obtains self-feedback, and requests further refinement. In each run, we first select an assessment question from a pre-constructed list of questions, as shown in Table 2. Customized for meta-review generation, this checklist covers the four most crucial aspects of meta-reviews. It can also be expanded and easily adapted to other complex text generation tasks. After prompting LLM with the assessment questions, we collect the refinement suggestions from the LLM's feedback. These refinement suggestions are further used as prompts for generating a revised version of the meta-review. The checklist questions are posed sequentially in one iterative run, with the number of iterations set as a hyper-parameter in CGI^2.

Our proposed approach offers two key benefits.

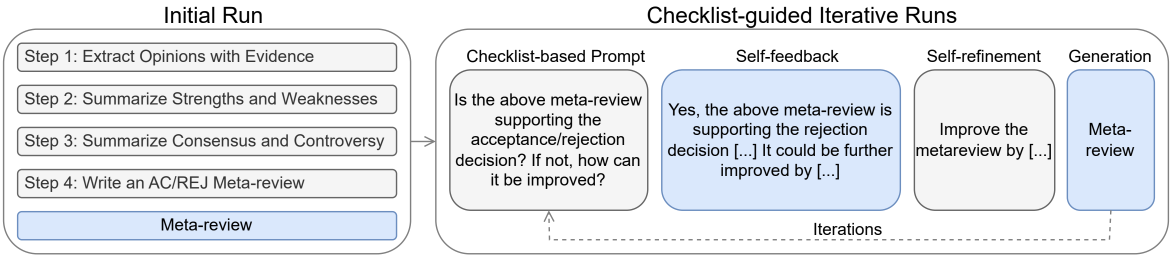

Figure 3: Our proposed CGI[2] framework operates through multiple iterations. In the initial iteration, the task is divided into four steps: (1) Review Opinion Extraction, (2) Strength and Weakness Synthesis, (3) Consensus and Controversy Analysis, and (4) Meta-review Drafting. For subsequent iterations, we present the black-box LLM with a query from a predefined list, acquire self-feedback, and request additional refinements.

1. Are the most important advantages and disadvantages discussed in the above meta-review? If not, how can it be improved?
2. Are the most important consensus and controversy discussed in the above meta-review? If not, how can it be improved?
3. Is the above meta-review contradicting reviewers' comments? If so, how can it be improved?
4. Is the above meta-review supporting the acceptance/rejection decision? If not, how can it be improved?

Table 2: The extensible and easily adaptable checklist for Meta-review Generation accesses the essential aspects of self-consistency, faithfulness, and active engagement in discussions.

First, it eliminates the need for external scoring functions that demand training data or human annotations. Second, it provides a general solution for employing GPT as a black box in complex text generation tasks.

6 Evaluation

Meta-review generation requires a system to accurately summarize opinions, highlight reviewer consensuses and controversies, offer judgments, and make recommendations. The task complexity thus requires an evaluation that is multifaceted and goes beyond n-gram similarity. However, current evaluation metrics for long text generation are inadequate for measuring the particular requirements of meta-review generation. To address this gap, we propose a comprehensive evaluation framework that combines standard evaluation metrics with LLM-based evaluation metrics.

6.1 Standard Metrics

We apply standard metrics in natural language generation to assess relevance, factual consistency, and semantic coherence. For relevance, ROUGE-L (Lin, 2004) quantifies the similarity between the generated and reference texts by calculating Longest Common Subsequence, while BERTScore (Zhang et al., 2020) offers a more nuanced relevance evaluation as it leverages the contextualized embeddings without relying on n-gram overlaps. For factual consistency, FACTCC (Kryscinski et al., 2019) checks whether a given claim in the generated text is consistent with the facts presented in the source document, while SummaC (Laban et al., 2021) utilizes sentence-level natural language inference models for inconsistency detection. DiscoScore (Zhao et al., 2022) presents six BERT-based model variants to measure discourse coherence. We average the scores from these six models as the coherence indicator.

6.2 LLM-based Metrics

The aforementioned methods do not evaluate discussion engagement or evidence-decision consistency. Some reference summaries may not include discussions or utilize evidence to substantiate decisions. To address this, we propose supplementary measures for this task that can be assessed and quantified using reference-free LLM-based metrics. We aim at assessing the following key aspects:

- Discussion Involvement: whether the meta-review discusses the paper's strengths and weaknesses, as well as agreements and disagreements among reviewers.
- Opinion Faithfulness: whether the meta-review contradicts reviewers' comments.
- Decision Consistency: whether the meta-review accurately reflects the final decisions.

Since our requirements cannot be described as simply as one word, we explore GPT-based evaluators other than GPTScore (Fu et al., 2023). G-EVAL (Liu et al., 2023) assesses the quality of NLG outputs by utilizing chain-of-thought (CoT)

G-EVAL

You will be given one metareview written for reviews by the committee on a paper. Your task is to rate the metareview on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria: Quality of Metareview (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby the metareview should be well-structured and well-organized. The metareview should always discuss the disadvantages and advantages of a paper and have a clear scope of the accept/reject decision. The metareview should have concrete evidence from the papers reviews and concrete comments as well.
Evaluation Steps:
1. Read the reviews carefully and identify the main topic and key points.
2. Read the metareview and compare it to the reviews. Check if the metareview covers the main topic, discusses advantages and disadvantages, if the most important advantages and disadvantages discussed in the above meta-review, if the most important advantages and disadvantages discussed in the above meta-review, if the most important consensus and controversy discussed in the above meta-review, if the above meta-review contradicting reviewers' comments, if the above meta-review supporting the acceptance/rejection decision, and if it presents them in a clear and logical order.
3. Assign a score for the quality of the meta-review on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.
Source Text: {Reviews} Metareview: {Meta-review} Evaluation Form (scores ONLY): - Quality of metareview :

Likert scale scoring with ChatGPT

Imagine you are a human annotator now. You will evaluate the quality of metareviews written for a conference by giving a mean value from 1 to 5 and no other explanation. Please follow these steps:
1. Carefully read the reviews, and be aware of the information it contains.
2. Read the proposed metareview.
3. Rate the summary on three dimensions: 'Discussion Involvement', 'Opinion Faithfulness' and 'Decision Consistency'. You should rate on a scale from 1 (worst) to 5 (best) and give me an average of these scores over all aspects from 1 to 5 calculated by the mean of all aspects.
Definitions are as follows:
(1) Discussion Involvement: Whether the meta-review discusses the paper's strengths and weaknesses, as well as agreements and disagreements among reviewers,
(2) Opinion Faithfulness: Whether the meta-review contradicts reviewers' comments,
(3) Decision Consistency: Whether the meta-review accurately reflects the final decisions.
Only generate the mean rating as a number on the likert scale, nothing else.

Figure 4: We customize the prompts in G-EVAL (Liu et al., 2023) and GPTLikert (Gao et al., 2023) for evaluating meta-review generation to assess discussion involvement, opinion faithfulness, and decision consistency.

and a form-filling paradigm and has shown a very high correlation with human-based judgments. G-EVAL uses carefully constructed instructions for GPT models to follow, which subsequently yields a rating on a Likert scale ranging from 1 to 5. Likert scale scoring with ChatGPT (GPTLikert), a human-like automatic evaluation method introduced in (Gao et al., 2023) that also outperforms many standard metrics in human correlation, follows a similar evaluation protocol. These methods have shown better human alignment on multiple text summarization tasks. We are the first to adapt these methods to meta-review generation by modifying the prompts as shown in Figure 4.

7 Experiments

7.1 Baselines

We compare our proposed CGI^2 method with methods in different paradigms. Results in Table 3 are average across three random runs.

Abstractive Methods. PlanSum(Amplayo et al., 2021b) uses a Condense-Abstract Framework, where reviews are condensed and used as input to an abstractive summarization model. OpinionDigest (Suhara et al., 2020) extracts opinions from input reviews and trains a seq2seq model that generates a summary from a set of these opinions. MeanSum (Chu and Liu, 2019) is an unsupervised multi-document abstractive summarizer that minimizes a combination of reconstruction and vector similarity losses. LED (Beltagy et al., 2020) is a Longformer variant supporting long document generative sequence-to-sequence tasks.

Extractive Methods. LexRank (Erkan and Radev, 2004) is an unsupervised extractive summarization method that selects sentences based on centrality scores calculated with graph-based sentence similarity. MemSum (Gu et al., 2022) models extractive summarization as a multi-step episodic Markov Decision Process of scoring and selecting sentences.

Prompting Methods. All prompting methods are initiated with the gpt-3.5-turbo model with a temperature of 0.7. 3Sent (Goyal et al., 2022) applies a simple prompt "Summary of document in 3 sentences". TCG (Bhaskar et al., 2022) explores a four-step generation pipeline involving topic classification, sentence grouping by topic, generating chunk-wise summary, and generating the final summary. We also explore In Context Learning (ICL) (Brown et al., 2020), where a highly rated meta-review alongside the reviews is given as part of a prompt to the model. This metareview is manually picked based on adherence to the checklist mentioned above and is chosen for its fulfillment of all the criteria that define a high-quality metareview. Vanilla uses "Generate a metareview" as the prompt. InstructPrompt provides more detailed instructions, which we show in the Appendix.

Models	ROUGE-L	BERTScore	FactCC	SummaC	DiscoScore	G-EVAL	GPTLikert
Human	-	-	0.538	0.368	0.740	0.731	0.607
Abstrative Methods							
PlanSum	**0.465**	0.785	0.608	0.533	0.911	0.731	0.608
OpinionDigest	0.124	0.838	0.612	0.575	0.862	0.762	0.618
MeanSum	0.132	0.827	0.559	0.464	0.900	0.767	0.622
LED	0.161	0.846	0.618	0.785	0.958	0.731	0.624
LED-finetuned	0.221	0.853	0.634	0.795	0.961	0.751	0.649
Extractive Methods							
LexRank	0.433	**0.881**	**0.729**	**0.937**	**1.256**	0.726	0.656
MemSum	0.337	0.827	0.683	0.825	0.989	0.711	0.628
Prompting Methods							
Vanilla	0.174	0.817	0.498	0.423	0.808	0.752	0.626
3Sent	0.109	0.783	0.562	0.503	0.667	0.758	0.661
InstructPrompt	0.208	0.823	0.543	0.449	0.862	0.751	0.646
TCG	0.189	0.847	0.544	0.466	0.895	0.761	0.632
ICL	0.192	0.847	0.578	0.470	0.871	0.756	0.612
CGI^2 (ours)	0.201	0.835	0.559	0.328	0.899	**0.768**	**0.673**
CGI^2 w/o Iterative Runs	0.118	0.830	0.536	0.332	0.849	0.732	0.629

Table 3: ROUGE-L and BERTScore assess semantic similarity with reference text. FactCC and SummaC detect factual consistency. DiscoScore measures coherence. G-EVAL and GPTLikert are GPT-based comprehensive evaluation measures for discussion involvement, opinion faithfulness, and decision consistency.

Model	Informativeness	Soundness	Self-consistency	Faithfulness
Human	0.71	0.68	0.67	-
LED-finetuned	0.56	0.46	0.21	0.73
LexRank	0.87	0.94	0.16	-
CGI^2 (ours)	**0.98**	**0.92**	**0.84**	**0.79**
CGI^2 w/o Iterative Runs	0.97	0.76	0.48	0.74

Table 4: Human annotation results on meta-reviews for 50 challenging papers from the test set.

7.2 Automatic Evaluation

Higher standard metric results indicate better summarization, but not necessarily better opinion summarization. ROUGE-L, BERTScore, SummaC, and DiscoScore do not consider the multifaceted nature of meta-review, which goes beyond summarization. Our method performs near average in BERTScore and SummaC and the highest in ROUGE-L and DiscoScore amongst the prompting baselines. When compared to extractive and abstractive methods, our method performs lower since some of them specifically account for maximizing semantic similarity.

Evaluators like G-Eval and GPTLikert favor specific dimensions given in their prompts. Our method shows promising results in both G-Eval and GPTLikert due to the carefully constructed and revised prompts. Most prompting methods also outperform extractive and abstractive methods.

Human meta-reviews in the dataset scored amongst the lowest in all categories, signifying the unreliability of some human-written meta-reviews and the need for the automatic writing auxiliary process. When comparing for semantic similarity, extractive methods outperform both abstractive and prompting methods with the exception of Plansum. This is due to the nature of content planning in Plansum which is very central to the task of meta-review generation.

7.3 Human Evaluation

We conduct a human annotation on 50 challenging boundary papers from the test set, which have average review scores on the borderline of acceptance. Five anonymous baseline outputs from Human, LED-finetuned, LexRank, CGI^2, and CGI^2 without iterative runs, are shown to three annotators. The annotators are asked to provide binary labels of informativeness, soundness, self-consistency, and faithfulness for each meta-review. Informativeness measures whether the meta-review involves both strength and weakness discussion. Soundness examines whether the meta-review provides evidence to support the discussed strength or weakness. Self-consistency indicates whether the recommendation decision is clearly written and consistent with the comments in the meta-review. Faithfulness evaluates whether the meta-review contains halluci-

Human

The paper unanimously receives positive rates thanks to strong motivations and interesting results. As the reviews show satisfaction on the authors' feedback, the final draft needs to respect it accordingly, for example, about the limitations of this research.

Vanilla

This paper could have benefited from a clearer and more concise presentation of the main findings. It is sometimes difficult to follow the paper's chain of thought, and the authors could have presented their results in a more accessible way. The experiments are limited to a small set of environments, and it would be interesting to see how the results generalize to other tasks. Additionally, it would be useful to have a deeper analysis of why iGPT does not perform well, as this could help inform future research on how to improve image-based pretraining.

CGI^2

After carefully considering the reviews provided, there is consensus that the paper presents an important and thought-provoking analysis of representation dynamics between pretraining and finetuning in the context of reinforcement learning. Both reviewers agree that the paper addresses an important question and presents insightful experiments. One reviewer highlights the experiments as a strong point in the paper, while the other reviewer raises a concern about the paper's tendency to present hypotheses without exploring them further.
Overall, the strengths of the paper outweigh its weaknesses, and it has the potential to make a significant contribution to the field of reinforcement learning. Therefore, I recommend accepting the paper with a minor revision to address the concern raised by the reviewer. Specifically, the paper should explore the hypotheses presented in more depth to provide a more comprehensive analysis. If the authors can address this concern, the paper can make a valuable contribution to the field of reinforcement learning.
In summary, I recommend accepting the paper with a minor revision.

Initial run

Based on the reviews, I recommend accepting the paper. The paper addresses an important and under-explored research question and presents a detailed and thought-provoking analysis of the pretraining-finetuning paradigm for Transformer models in the context of reinforcement learning. The paper also identifies relevant techniques to perform analysis of representation similarity and training dynamics. While there are some weaknesses pointed out, such as the need for further exploration of some hypotheses and occasional difficulty in following the writing and presentation, the strengths of the paper outweigh these concerns. Therefore, I recommend accepting the paper with minor revisions to address the weaknesses pointed out by the reviewers.

Figure 5: We show the meta-reviews from human, vanilla, CGI^2, and CGI^2 without iterative runs for the same paper. The yellow background indicates hallucinated content. The green background indicates redundant content.

nations. We assume Human and the extractive LexRank have perfectly faithful summaries.

Results shown in Table 4 validate the effectiveness of our proposed method. The extractive method (LexRank) is easily biased toward one reviewer, involving no discussion nor decision, but having no hallucination problems. The abstractive method (LED-finetuned) learns to copy the sentences in the input and form a short meta-review with little discussion and sometimes internal hallucinations or repetitiveness. Our prompting-based method presents less hallucination with the evidence requirements in designed prompts. Compared to human-written meta-reviews, all automatic methods are less capable of generating in-depth analysis, which calls for knowledge enhancement.

We also observe that hallucinations in LLM are more likely to happen in summarizing consensus and controversy, which requires information integration. In contrast, hallucinations in the extractive-alike abstractive method are more likely to be triggered by generating some general comments. Hallucination detection in scientific opinion summarization remains an opening problem.

7.4 Case Study

Figure 5 presents the meta-reviews from human, vanilla, CGI^2, and CGI^2 without iterative runs for a random paper[3].

From the qualitative results, we have the following observations: (1) The hallucination problem is alleviated in CGI^2 because the model is constantly asked for evidence. (2) The language style of always providing a summary at the end brings redundancy in CGI^2. (3) The vanilla prompting baseline usually does not make recommendations and involve discussion, as the model fails to fully understand the complex task requirement. (4) Iterative refinement sometimes improves the concreteness of opinion discussion. However, there are two problems with the iterative refinements. First, the suggestions provided by the large language model are usually generic and less useful for further refinement. Second, more self-refinement iterations bring heavier forgetfulness for the initial instructions on opinion extraction and discussion.

8 Conclusions and Future Work

In this paper, we introduce the task of scientific opinion summarization, where research paper reviews are synthesized into meta-reviews. To facilitate this task, we introduce a new ORSUM dataset, an evaluation framework, and a Checklist-Guided Iterative Introspection approach. We conduct an empirical analysis using methods in different paradigms. We conclude that human-written summaries do not always accommodate all necessary criteria, and the combination of task decomposition and iterative self-refinement shows promising discussion involvement ability and can be applied to other complex text generation using black-box LLM.

Direct extensions of this work include the incorporation of author rebuttals into the input data to enhance the model's ability to generate more balanced meta-reviews and the extension of a more comprehensive checklist.

[3]https://openreview.net/forum?id=9GXoMs__ckJ

Limitations

This work on scientific opinion summarization has limitations in terms of data scope and task configuration. As the dataset is collected from OpenReview, the majority of meta-reviews are in the Machine Learning area, and many papers have been accepted. Conclusions drawn from this data distribution might not be applicable to datasets in other domains. Furthermore, to simplify the task setting, author rebuttals have not been included as input, which may also constrain the extent of discussion engagement in generating meta-reviews.

Ethics Statement

We acknowledge the following potential ethical concerns that may arise. First, the meta-reviews generated by LLMs may contain hallucinations, which may lead to misunderstandings of the original research paper or reviewers' opinions. Therefore, users should be cautious when using system-generated meta-reviews for recommendation decisions. Second, the use of black-box LLMs for meta-review generation may raise concerns about the transparency of the decision process. Though our method improves explainability by prompting an LLM to provide supporting evidence for the recommendation decision, the evidence may not perfectly reflect the decision-making process. Third, the dataset used in this study mainly focuses on the machine learning area, which might introduce biases to the recommendation decisions. Hence, it is critical to consider these biases when applying our method to generate meta-reviews for research papers in other domains.

References

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

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021b. Unsupervised opinion summarization with content planning. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 12489–12497. AAAI Press.

Reinald Kim Amplayo and Mirella Lapata. 2019. Informative and controllable opinion summarization. CoRR, abs/1909.02322.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 1934–1945. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2020. Extractive opinion summarization in quantized transformer spaces. CoRR, abs/2012.04443.

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

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 3675–3686. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. CoRR, abs/2004.05150.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2022. Zero-shot opinion summarization with GPT-3. CoRR, abs/2211.15914.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 1653–1656. ACM.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 4119–4135. Association for Computational Linguistics.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 5151–5169. Association for Computational Linguistics.

Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. In Findings of the Association for Computational Linguistics: NAACL

2022, Seattle, WA, United States, July 10-15, 2022, pages 1509–1523. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Hou Pong Chan, Wang Chen, and Irwin King. 2020. A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 1191–1200. ACM.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of Findings of ACL, pages 5062–5074. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: argument pair extraction from peer review and rebuttal via multi-task learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 7000–7011. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 1209–1225. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 1223–1232. PMLR.

Zhongfen Deng, Hao Peng, Congying Xia, Jianxin Li, Lifang He, and Philip S. Yu. 2020. Hierarchical bi-directional self-attention networks for paper review rating recommendation. In Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pages 6302–6314. International Committee on Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

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

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

Korbinian Friedl, Georgios Rizos, Lukas Stappen, Madina Hasan, Lucia Specia, Thomas Hain, and Björn W. Schuller. 2021. Uncertainty aware review hallucination for science article classification. In Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of Findings of ACL, pages 5004–5009. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. CoRR, abs/2302.04166.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. CoRR, abs/2209.12356.

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

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA, pages 1621–1624. AAAI Press.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex aggregation for opinion summarization. CoRR, abs/2104.01371.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance. Transactions of the Association for Computational Linguistics, 9:945–961.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. CoRR, abs/1910.12840.

Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. 2021. A deep neural architecture for decision-aware meta-review generation. In ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021, Champaign, IL, USA, September 27-30, 2021, pages 222–225. IEEE.

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

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, pages 345–354. ACM.

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. CoRR, abs/2303.16634.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

OpenAI. 2021. Chatgpt-3.5-turbo. `https://openai.com/research/`. Accessed: 2023-05-20.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 2521–2535. Association for Computational Linguistics.

Ming Shen, Jie Ma, Shuai Wang, Yogarshi Vyas, Kalpit Dixit, Miguel Ballesteros, and Yassine Benajiba. 2023. Simple yet effective synthetic dataset construction for unsupervised opinion summarization. CoRR, abs/2303.11660.

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, ACL 2020, Online, July 5-10, 2020, pages 5789–5798. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 47–57. The Association for Computational Linguistics.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020, pages 384–397. Association for Computational Linguistics.

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, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, pages 516–528. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Wei Zhao, Michael Strube, and Steffen Eger. 2022. Discoscore: Evaluating text generation with BERT and discourse coherence. CoRR, abs/2201.11176.

Appendix

A Opinion Summarization Datasets

The Rotten Tomatoes (RT) dataset (Wang and Ling, 2016) consists of movie critics and their editor-written one-sentence opinion consensus. Copycat (Brazinskas et al., 2020b) and OPOSUM (Angelidis and Lapata, 2018) annotate small reference evaluation sets for Amazon products with Amazon Mechanical Turk (AMT). Another human-annotated set (Chu and Liu, 2019) from Yelp reviews has 200 AMT-annotated summaries. DENOISESUM (Amplayo and Lapata, 2020) creates a synthetic dataset from RT (Wang and Ling, 2016) and Yelp (Chu and Liu, 2019) by sampling a review as a candidate summary and generating noisy versions as its pseudo-review inputs, where reviews not reaching consensus will be treated as noise. PLANSUM (Amplayo et al., 2021b) is another synthetic dataset from RT (Wang and Ling, 2016), Yelp (Chu and Liu, 2019), and Amazon (Brazinskas et al., 2020b) created by sampling pseudo-reviews from a Dirichlet distribution parametrized by a content planner. SPACE (Angelidis et al., 2021) creates a collection of human-written general summaries and aspect summaries for 50 hotels.

B ORSUM Composition Annotation

We select 100 meta-reviews to conduct a human annotation for meta-review composition. We draw one meta-review from each venue and randomly select the others from the rest of the training set.

We ask three annotators to label the meta-review composition in two dimensions: whether the meta-review contains a detailed discussion of the paper's strengths and weaknesses, and whether the meta-review includes specific comments on the agreements and disagreements among the reviews. The scores range from 0 to 2, with the following interpretations: 0 indicates that the meta-review does not address the discussion at all. 1 signifies that the meta-review incorporates the discussion but lacks concrete evidence. 2 denotes that the meta-review involves a detailed discussion. For example, "The three reviewers agreed that the contribution is relevant to the workshop and presents a solid work. " is assigned a score of 1 in both dimensions because, while it refers to the discussion, the comment remains generic. The annotation process is conducted at the sentence level. If a meta-review contains a sentence with a score of 2, the entire meta-review is labeled with a score of 2.

C Implementation Details

Due to the input length constraint, each review is truncated to 300 tokens. For iterative runs in CGI^2, given the number of instructions, the reviews are deleted from the appended messages, and only discussion of these reviews with the respective evidence and initial metareview are passed forward. Similar truncation is done in the prompting-based evaluators.

For LED we use the LEDforConditionalGeneration model from Huggingface. For MeanSum and OpinionDigest, we use their provided pretrained models. We train the content induction model of Plansum on ORSUM. In CGI^2, we set the number of iterations to 1. We show the used prompts in Table 5.

D Examples of Generated Meta-Reviews

We show three generated examples in Table 6.

Models	**Prompts**
Vanilla	Generate a Metareview
3sent	Generate a summary of document in 3 sentences.
InstructPrompt	Imagine you are a human metareviewer now. You will write metareviews for a conference. Please follow these steps: 1. Carefully read the reviews, and be aware of the information it contains. 2. Generate a metareview based on three dimensions: 'Discussion Involvement', 'Opinion Faithfulness' and 'Decision Consistency'. Definitions are as follows: (1) Discussion Involvement: Discuss the paper's strengths and weaknesses, as well as agreements and disagreements among reviewers, (2) Opinion Faithfulness: Do not contradict reviewers' comments, (3) Decision Consistency: Accurately reflect the final decisions.
TCG	Describe the topic of each sentence in one word. Summarize what reviewers said of the paper. Summarize the summaries of the reviews
ICL	Given a pair of reviews and a metareview as an example, Generate a metareview based on given reviews. {example}
CGI^2 (ours)	From the sentiments and aspects discussed in the reviews, what are the key strengths and weaknesses of this paper? Please cite corresponding reviewers and evidence. Identify the points of agreement and disagreement among the reviewers. Please include the corresponding reviewers and evidence. Considering the key sentiments from the reviews, the identified strengths and weaknesses, and the consensus and controversy among the reviewers, write a metareview with decision of acceptance/rejection. Are the most important advantages and disadvantages discussed in the above meta-review? Are the most important consensus and controversy discussed in the above meta-review? Is the above meta-review contradicting reviewers' comments? Is the above meta-review supporting the rejection decision? If not, how can it be improved? using this discussion, write an accurate <200 words metareview with decision of acceptance/rejection.
CGI^2 ablation	From the sentiments and aspects discussed in the reviews, what are the key strengths and weaknesses of this paper? Please cite corresponding reviewers and evidence. Identify the points of agreement and disagreement among the reviewers. Please include the corresponding reviewers and evidence. Considering the key sentiments from the reviews, the identified strengths and weaknesses, and the consensus and controversy among the reviewers, write a metareview with decision of acceptance/rejection.

Table 5: Prompts we use in all prompting-based methods.

After careful consideration of the reviews and feedback, the proposed paper presents a novel method for constructing variational approximations using SMC with partial rejection control and provides empirical validation of the approach on a variational RNN. The paper also provides clear explanations of the technical details and limitations of the approach. While the notation used in the paper can be cumbersome, this is a common issue in papers on SMC algorithms. However, there are also some weaknesses identified by the reviewers, including the complexity of the new bound and its scalability with K, the need for more clarity on the connection between the variance of the estimator for the normalizing constant obtained from particle filters and the tightness of the variational bound, and confusion over the use of Bernoulli factories and dice enterprise in the paper. Despite these weaknesses, the overall consensus among the reviewers is that the paper is a valuable contribution to the field and the proposed method has potential for further development. Therefore, the paper is recommended for acceptance with minor revisions.
The paper proposes a deep reinforcement learning model for portfolio optimization that harnesses both cross-asset dependencies and time dependencies. The proposed approach is evaluated on three sets of data from the Canadian and US stock markets, and the results show that the proposed approach outperformed baselines in terms of different metrics such as annual return and Sharpe ratio. The paper presents a clear formulation of the portfolio optimization problem and a reasonable solution to deal with permutation invariance across assets. The proposed permutation invariant architecture is also effective and has potential applications beyond portfolio management. Additionally, the theoretical characterizations presented in the paper are appreciated. However, there are some concerns about the clarity of technical details, such as the description of the architecture and the RL algorithm used to optimize the policy. There are also questions about the motivation for using the Corr layer instead of popular methods such as graph neural networks or transformers. While the contribution of dealing with permutation invariance is recognized, some reviewers note that this property is not new.
The proposed visual-inertial odometry (VIO) system is a novel and effective method that achieves competitive results, according to the four reviewers' feedback. The paper presents a clear and technically sound methodology that uses differentiable Unscented Kalman Filter (UKF) to learn the covariance matrices in an end-to-end manner. The authors' approach is refreshing as it is a learning-based VIO work that is not based on deep networks. However, there are some concerns raised by the reviewers. One reviewer pointed out the lack of empirical evaluation of the model interpretability, while the others highlighted the limited evaluation based mainly on the KITTI dataset and the lack of analysis on the sufficient distance assumption. Despite these weaknesses, the paper is well-written and presents significant novelties and contributions. Therefore, the paper is recommended for acceptance with minor revisions addressing the concerns raised by the reviewers. The authors should consider providing more empirical evaluation of the model interpretability.

Table 6: Examples of the meta-reviews generated by our proposed CGI^2 method.