Aspect-Aware Decomposition for Opinion Summarization

Anonymous ACL submission

Abstract

Opinion summarization plays a key role in deriving meaningful insights from large-scale online reviews. To make this process more explainable and grounded, we propose a modular approach guided by review aspects (e.g., cleanliness for hotel reviews) which separates the tasks of aspect identification, opinion consolidation, and meta-review synthesis, enabling greater transparency and ease of inspection. We conduct extensive experiments across datasets representing scientific research, business, and product domains. Results show that our method generates more grounded summaries compared to strong baseline models, as verified through automated and human evaluations. Additionally, our modular approach, which incorporates reasoning based on review aspects, produces more informative intermediate outputs than knowledge-agnostic decomposed prompting. These intermediate outputs can also effectively support humans in summarizing opinions from large volumes of reviews.

1 Introduction

Reviews are omnipresent in the digital world, providing invaluable insights into products (Bražinskas et al., 2021), businesses (Angelidis et al., 2021), even scientific articles (Li et al., 2023). Automatic opinion summarization aims to *aggregate* a large and diverse set of reviews about a particular *entity* (e.g., hotel) into a single easy-to-read *meta-review* (or summary). A good meta-review should accurately reflect the balance of opinions in the source reviews and speak to the entity's most important *aspects* (e.g., *Cleanliness*, *Service*, *Location*). A useful meta-review should also present some *evidence* justifying its content.

Opinion summarization has distinct characteristics that set it apart from other summarization tasks. Firstly, it cannot rely on reference summaries for training, as human-written meta-reviews are not generally available (e.g., across entities and domains) and can be difficult to crowdsource (e.g., for entities represented by thousands of reviews). Secondly, methods need to be flexible with respect to the scope of the output. Users may wish to read a general meta-review covering *all* aspects related to the entity of interest, or a more targeted one focusing on *specific* aspects. Finally, given the subjective nature of the summarization task, systems should offer some evidence to justify their output.

Prior approaches to generating meta-reviews broadly fall into three categories. *Extractive* methods create summaries by selecting a few representative sentences from source reviews (Angelidis et al., 2021; Basu Roy Chowdhury et al., 2022). While these approaches are scalable and inherently attributable, the summaries tend to be overly detailed and lack coherence. *Abstractive* methods rely on neural language models to generate fluent and coherent meta-reviews with novel language (Frermann and Klementiev, 2019; Chu and Liu, 2019; Coavoux et al., 2019; Bražinskas et al., 2020; Amplayo et al., 2021a,b; Iso et al., 2021; Bražinskas et al., 2021; Cattan et al., 2023). However, most abstractive approaches are neither attributable nor controllable due to the black-box nature of end-to-end modeling and face issues with input length (e.g., due to context window limits).

Hybrid approaches (Hosking et al., 2023, 2024; Li et al., 2024) cluster sentences according to some criterion (e.g., similarity or sentiment) and then generate summaries (e.g., using a language model) based on the clusters containing the most *popular* opinions. The summaries are fluent and attributable since the output is associated with evidential clusters, but the quality of the clusters can vary, requiring additional post-processing and it is not immediately clear how to consider well-justified opinions rather than the most popular ones.

In this paper, we propose to decompose opinion summarization into simpler sub-tasks that can be executed by prompt-based Large Language Mod-

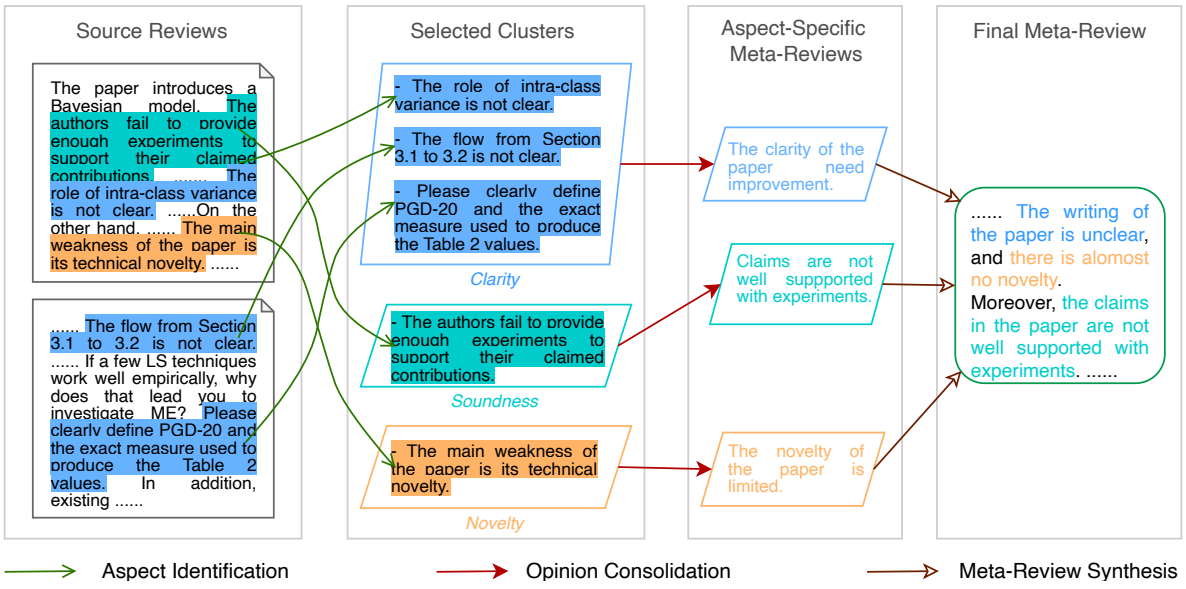

Figure 1: High-level overview of our decomposition for opinion summarization using an example from the scientific domain with three aspects (*Clarity*, *Soundness*, and *Novelty*). The modules *Aspect Identification*, *Opinion Consolidation*, and *Meta-Review Synthesis* are instantiated with prompt-based LLMs and operate in sequence. The output of *Aspect Identification* serves as input to *Opinion consolidation* and *Meta-Review synthesis* aggregates opinions found in aspect-specific meta-reviews. All prompts and inputs/outputs are in natural language.

els (LLMs) dedicated to these sub-tasks. Our approach is inspired by recent applications of chain-of-thought prompting (Wei et al., 2022) and its variants (Khot et al., 2023; Zhou et al., 2023), which address reasoning problems by decomposing complex tasks into a sequence of simpler sub-problems which are solved sequentially. Our decomposition consists of three high-level modules, namely *Aspect Identification*, *Opinion Consolidation*, and *Meta-Review Synthesis*. Intuitively, we first identify text fragments in the input reviews discussing aspects pertaining to the entity and domain in question; next we create meta-reviews for *each* aspect, and finally we generate a global meta-review for *all* aspects (see Figure 1). Our approach eschews problems relating to the scale of the input, since reviews can be processed in parallel to identify the aspects. It also avoids problems with clusters being diffuse or irrelevant since we leverage domain specific aspect definitions (as part of the prompt) to obtain interpretable clusters. Finally, our decomposition is controllable, and evidence-based, as the output of each module can be traced back to its input. Our contributions can be summarized as follows:

- We propose a decomposition of opinion summarization into three modules which can be instantiated with LLMs using zero-shot prompting. Our decomposition is domain agnostic, controllable, and evidence-based.
- Extensive experiments on three datasets from different domains demonstrate that our aspect-informed approach produces more grounded meta-reviews than strong baselines in terms of automatic and human evaluation.
- Compared to automatic prompt decomposition methods (Khot et al., 2023), we show that task-aware decomposition yields more useful reasoning chains and intermediate outputs, which could assist humans with summarizing reviews.

2 Related Work

Our work focuses on abstractive opinion summarization that aims to generate fluent and coherent summaries with novel language (Bražinskas et al., 2021; Li et al., 2023). This task has been explored in different domains, such as summarizing reviews of products, businesses, and scientific articles (Chu and Liu, 2019; Bražinskas et al., 2021; Li et al., 2023; Hosking et al., 2024). Previous abstractive methods can only process a limited number of source reviews, and lack transparency in their decision-making process due to their end-to-end nature. Hybrid approaches implement pipelines with transparent intermediate outputs, however, they are aspect agnostic, focusing on how to organize or annotate the input for downstream processing.

For example, Hosking et al. (2024) propose a method that represents sentences from reviews as paths through a learned discrete hierarchy, and then use LLMs to generate output sentences based on frequent paths retrieved from this hierarchy. Their retrieval module relies heavily on majority voting, which is less effective in domains where minority but well-argued opinions are valuable, such as in scientific reviews. Li et al. (2024) generate meta-reviews exclusively for the scientific domain, framing the task as a form of sentiment summarization. Their method extracts sentiments from reviews taking into account how these are structured (e.g., into opinions on *Novelty* and *Soundness*).

A few approaches take aspects into account, and are thus able to produce both general and aspect-specific opinion summaries. Angelidis et al. (2021) achieve this by clustering opinions through a discrete latent variable model and extracting sentences based on popular aspects or a particular aspect. Other work fine-tunes pre-trained models on synthetic data enhanced with aspect annotations which can be used to control output summaries at inference time (Amplayo et al., 2021a). Our own work delegates the task of aspect identification to prompt engineering, demonstrating that LLMs can reliably extract aspects given an input review and aspect definitions without additional training. We make no assumptions regarding the structure of the summaries, and how aspects should be presented in them — we assume these can be tailored using appropriate instructions to suit specific users and domains.

Our work relates to recent efforts aiming to improve the in-context learning performance of LLMs through intermediate reasoning chains (Wei et al., 2022; Yao et al., 2023; Khot et al., 2023). Previous approaches focus primarily on mathematical or symbolic reasoning, while intermediate reasoning for complex writing tasks such as opinion summarization remains under-explored (Li et al., 2024). Decomposed prompting (Khot et al., 2023) is a recent approach to solving complex tasks using (few-shot) LLMs by predicting both the task decomposition into modules and the modules themselves. We adapt to our task a well-known decomposition of multi-document summarization (Barzilay and McKeown, 2005; Radev and McKeown, 1998; Lebanoff et al., 2020; Slobodkin et al., 2024; Krishna et al., 2021; Li et al., 2024) into three modules, namely content selection, content consolidation (or fusion), and output generation. We also empirically find that automatic knowledge-agnostic task decomposition is inferior, at least in the context of opinion summarization.

3 Task Decomposition

Let $\mathcal{C}$ denote a corpus of reviews on entities $\{e_1, e_2, \dots\}$ from a domain d, for example, hotels or scientific articles. Reviews may discuss a number of relevant aspects $A_d = \{a_1, a_2, \dots\}$, like *Clarity* or *Soundness*, For each entity e_i, our task is to generate the meta-review $\widehat{y}_i$ by synthesizing opinions from a set of source reviews $R_i = \{r_1, r_2, \dots\}$ covering *all* attested aspects A_d. We decompose the task into three modules, namely *Aspect Identification*, *Opinion Consolidation*, and *Meta-Review Synthesis*. We present the inner workings of each module in Figure 1 with an example from the scientific domain. Due to the limited availability of training data, we implement our modules using an unsupervised approach, leveraging zero-shot prompting of LLMs and their instruction following and generation capabilities.[1]

Aspect Identification As not all content in the source reviews is relevant for generating meta-reviews, opinion summarization models must be able to isolate critical information in the input. The first module, *Aspect Identification*, selects text fragments of variable lengths from source reviews discussing any review aspect. Specifically, for reviewed entity e_i, our module identifies text fragments for aspect a_j from the source reviews R_i. The module essentially partitions text fragments into aspect-specific clusters $C_{i,j} = \{f_1, f_2, \dots\}$, where fragments f_m can originate from any source review in R_i. For example, in Figure 1, the module identifies fragments in scientific reviews for the aspects *Clarity*, *Soundness*, and *Novelty*. We implement this module with zero-shot LLM prompting. Our prompt template is shown in Appendix A (Figure 3) and can be modified for different aspects and domains.

Opinion Consolidation As shown in Figure 1, the output of the first module consists of clusters of text fragments, each discussing a specific aspect. Depending on the domain, these clusters can have a lot of redundancy, often repeating the same opinion. Our second module, *Opinion Consolidation*, aggre-

[1] It is worth noting that our prompts could be further improved, however, we leave prompt optimization to future work.

gates opinions into aspect-specific meta-reviews. We essentially adopt a divide-and-conquer strategy, since generating meta-reviews from aspect-specific clusters is significantly easier than producing an entire summary from reviews containing mixed aspects. Specifically, taking as input cluster $C_{i,j}$, the module generates meta-review $o_{i,j}$ for aspect a_j. As we do not have training data for these intermediate summaries, we also implement this module with zero-shot prompting.[2] Our template (shown in the Appendix, Figure 4) instructs LLMs to integrate opinions (i.e., text fragments) from a specific cluster. For example, in Figure 1 the three sentences in the *Clarity* cluster are aggregated into "*The clarity of the paper needs improvement*".

Meta-Review Synthesis After obtaining all aspect-specific summaries $O_i = \{o_{i,1}, o_{i,2}, \ldots\}$, our last module generates the final meta-review $\widehat{y}_i$ for entity e_i; it combines the opinions mentioned in the individual summaries into a fluent and coherent overall summary. An example is given in Figure 1 where the meta-review focuses on the aspects of *Clarity*, *Soundness*, and *Novelty*. Again, this module leverages the generation capabilities of LLMs, and is instantiated via zero-shot prompting. Our template (given in the Appendix, Figure 5) asks the LLM to write a concise meta-review which summarizes the provided opinions and covers all mentioned aspects.

4 Experimental Setup

We showcase the versatility of our approach on different domains. In this section, we describe the datasets used in our experiments, discuss implementation details and comparison baselines, and explain how we evaluate performance with automatic metrics.

Datasets We conducted experiments on three domains, product reviews for sports shoes, business reviews for hotels, and scientific reviews for research articles. For business reviews, we use SPACE, an opinion summarization dataset constructed by Angelidis et al. (2021). For product reviews, we use the sports shoes subset from AmaSum (Bražinskas et al., 2021). For scientific reviews, we use PeerSum (Li et al., 2023) and also the human annotations of review aspects from Li et al. (2024). Statistics for these datasets are shown in Table 1.

Dataset	#Train/ Dev/Test	#Reviews	SourceL	MetaL	#Aspects
PeerSum	22,420/50/100	14.9	5,146	156.1	5
AmaSum	25,203/50/50	381.8	14,495	94.8	10
SPACE	0/25/25	100	14,439	75.7	6

Table 1: Statistics of our experimental datasets. #Train/Dev/Test refer to the number of training, development, and test instances, respectively; #Reviews is the average number of reviews per entity; SourceL refers to the total length of the source reviews (when concatenated) and MetaL to the average meta-review length; #Aspects is the number of aspects covered in each dataset. For AmaSum, the statistics are for the sports shoes subset.

SPACE (Angelidis et al., 2021) consists of hotel reviews from TripAdvisor, with 100 reviews per entity, as well as reference meta-reviews of customer experiences created by annotators. The dataset covers six aspects for hotels, which we adopt in our experiments, namely *Building*, *Cleanliness*, *Food*, *Location*, *Rooms*, and *Service*. AmaSum contains meta-reviews for a variety of Amazon products, with reference summaries collated from professional review platforms. We only use the sports shoes subset curated from the RunRepeat platform which covers the aspects: *Breathability*, *Durability*, *Weight*, *Cushioning*, *Stability*, *Flexibility*, *Traction*, *Size and Fit*, *Comfort*, and *Misc*. PeerSum (Li et al., 2024) contains reviews for scientific articles and corresponding meta-reviews from OpenReview focusing on the aspects of *Novelty*, *Soundness*, *Clarity*, *Advancement*, and *Compliance*. Detailed definitions for all aspects (SPACE, AmaSum, and PeerSum) are given in the Appendix B–D.

Model Comparisons We implement our modular approach with different backbone LLMs, including closed- and open-source models. Since the modules need to have reasonable language generation and instruction following capabilities, we conduct experiments with gpt-4o-2024-05-13[3] from OpenAI, and Llama-3.1-70B-Instruct[4] and Llama-3.1-8B-Instruct[5] from Meta.[6] The prompts used in our experiments are provided in Appendix B–D.

We compare our approach with representative prompting and fine-tuning baselines. We implement two strong prompting approaches which do

[2] Some aspects may not have corresponding text fragments in the source reviews, as they do not always cover every aspect.

[3] https://platform.openai.com/docs/models/gpt-4o

[4] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] All models used in our experiments are instruction-tuned.

Models	Coverage↑	G-Eval↑	AlignScore-R/M↑	Rouge↑
Sentiment CoT-GPT-4o (Li et al., 2024)	0.96	0.75	0.72/0.08	**23.47**
FT-Llama 8B (Touvron et al., 2023)	0.87	0.60	0.33/0.06	20.60
Aspect-aware decomposition-GPT-4o (ours)	0.95	0.76	0.68/0.06	20.78
Automatic decomposition-Llama 8B (Khot et al., 2023)	0.58	0.20	0.36/0.03	11.98
Chunk-wise decomposition-Llama 8B (Khot et al., 2023)	0.79	0.65	0.65/0.03	21.19
Naive aspect-aware prompting-Llama 8B (Radford et al., 2019)	0.72	0.62	0.70/0.06	16.93
Aspect-aware decomposition-Llama 8B (ours)	0.90	0.66	0.71/0.07	21.12
Automatic decomposition-Llama 70B (Khot et al., 2023)	0.59	0.31	0.51/0.03	12.0
Chunk-wise decomposition-Llama 70B (Khot et al., 2023)	0.84	0.72	0.65/0.06	21.80
Naive aspect-aware prompting-Llama 70B (Radford et al., 2019)	0.72	0.62	0.70/0.07	16.82
Aspect-aware decomposition-Llama 70B (ours)	**0.97**	**0.76**	**0.76/0.09**	22.58

Table 2: Results on scientific **reviews of research articles**. The first section of the table presents results for GPT-4o and state-of-the-art models. The second section has results for Llama-8B, and the third one for Llama 70B. Underlined scores denote best in section per metric while bold scores denote best overall. AlignScore-R calculates AlignScore against source reviews, while AlignScore-M is computed against reference meta-reviews.

not take aspect information into account: *automatic decomposition* breaks down complex reasoning tasks into simpler ones (Khot et al., 2023) by automatically predicting the decomposition and the modules, while *chunk-wise decomposition* (Khot et al., 2023) recursively summarizes the input reviews chunk-by-chunk with prompting.[7] We also compare against the *naive aspect-aware prompting* which does not perform task decomposition but is aspect-aware (Radford et al., 2019). For fine-tuning, we conduct experiments on decoder-only LLMs. Due to computational limitations, we present fine-tuning results only with Llama-3.1-8B[8] on all three datasets. Moreover, we also include generations from state-of-the-art approaches for the datasets (see more details in Appendix E).

Automatic Evaluation Metrics We evaluate the quality of generated meta-reviews in terms of *aspect coverage* and *faithfulness* (against source reviews). Aspect coverage measures how well the generated meta-review for entity e_i captures the aspects discussed in the source reviews. Specifically, we compute the F_1 between the set of aspects present in the generated meta-review and those in the source reviews. We recognize aspects automatically by running our Aspect Identification module (see Section 3) on the system input and output. Opinion faithfulness measures how well opinions in generated meta-reviews are supported by the source reviews. Specifically, we use G-Eval (Liu et al., 2023), a prompting-based evaluation[9] metric, and AlignScore (Zha et al., 2023)[10], a fine-tuned evaluation metric based on information alignment between two arbitrary text pieces. We use the large version of the pre-trained backbone for AlignScore, and we set *nli_sp* as our evaluation mode. We also report Rouge F1 (Lin and Hovy, 2003), as a measure of overall summary quality.

5 Results and Analysis

We perform experiments on datasets covering multiple domains, comparing meta-reviews generated by our approach with those from strong baselines and state-of-the-art approaches. We further evaluate the intermediate outputs obtained from our modules against human annotations and conduct ablations to examine the extent to which individual modules contribute to the summarization task. Finally, in addition to automatic evaluation we conduct human evaluation based on pair-wise system comparisons and intermediate outputs.

Aspect-aware decomposition leads to better aspect coverage and opinion faithfulness. Our results using automatic evaluation metrics are summarized in Table 2 (scientific articles), Table 3 (shoes), and Table 4 (hotels).[11] Across domains we find that our modular approach with GPT-4o or Llama-3.1-70B delivers the highest coverage of review aspects. Our approach with GPT-4o is also better than comparison systems in terms of opinion

[7] The input is chunked based on document boundaries. For PeerSum each review is a chunk, while for AmaSum and SPACE chunks correspond to 20% of the source documents.

[8] https://huggingface.co/meta-llama/Llama-3.1-8B

[9] Our prompts are provided in Appendix F.

[10] https://github.com/yuh-zha/AlignScore/tree/main

[11] We run inference three times, with different random seeds and report average performance.

Models	Coverage↑	G-Eval↑	AlignScore-R/M↑	Rouge↑
HIRO-abs (Hosking et al., 2024)	0.54	0.35	0.78/0.13	14.90
FT-Llama 8B (Touvron et al., 2023)	0.45	0.12	0.43/0.16	9.90
Aspect-aware decomposition-GPT-4o (ours)	**0.86**	0.87	**0.79/0.17**	16.10
Automatic decomposition-Llama 8B (Khot et al., 2023)	0.39	0.11	0.47/0.13	9.23
Chunk-wise decomposition-Llama 8B (Khot et al., 2023)	0.58	0.80	0.66/0.08	**16.59**
Naive aspect-aware prompting-Llama 8B (Radford et al., 2019)	0.54	0.29	0.50/0.07	8.80
Aspect-aware decomposition-Llama 8B (ours)	0.77	0.78	0.69/0.09	16.44
Automatic decomposition-Llama 70B (Khot et al., 2023)	0.31	0.28	0.68/0.14	7.74
Chunk-wise decomposition-Llama 70B (Khot et al., 2023)	0.57	**0.88**	0.54/0.07	15.28
Naive aspect-aware prompting-Llama 70B (Radford et al., 2019)	0.49	0.48	0.60/0.09	7.35
Aspect-aware decomposition-Llama 70B (ours)	0.83	0.86	0.74/0.16	16.40

Table 3: Results on product **reviews of sports shoes**. The first section of the table presents results for GPT-4o and state-of-the-art models. The second section has results for Llama-8B, and the third one for Llama 70B. Underlined scores denote best in section per metric while bold scores denote best overall. AlignScore-R calculates AlignScore against source reviews, while AlignScore-M is computed against reference meta-reviews.

Models	Coverage↑	G-Eval↑	AlignScore-R/M↑	Rouge↑
HIRO-abs (Hosking et al., 2024)	0.87	0.62	**0.83/0.24**	**26.50**
Aspect-aware decomposition-GPT-4o (ours)	**1.00**	**0.90**	0.81/0.10	21.38
Automatic decomposition-Llama 8B (Khot et al., 2023)	0.65	0.07	0.55/0.15	13.80
Chunk-wise decomposition-Llama 8B (Khot et al., 2023)	0.94	0.80	0.65/0.14	22.9
Naive aspect-aware prompting-Llama 8B (Radford et al., 2019)	0.55	0.06	0.34/0.18	10.30
Aspect-aware decomposition-Llama 8B (ours)	0.97	0.81	0.70/0.10	22.05
Automatic decomposition-Llama 70B (Khot et al., 2023)	0.63	0.38	0.70/0.22	10.0
Chunk-wise decomposition-Llama 70B (Khot et al., 2023)	0.93	0.84	0.65/0.01	22.02
Naive aspect-aware prompting-Llama 70B (Radford et al., 2019)	0.37	0.34	0.44/0.22	5.00
Aspect-aware decomposition-Llama 70B (ours)	0.99	0.88	0.79/0.11	23.46

Table 4: Results on business **reviews of hotels**. The first section of the table presents results for GPT-4o and state-of-the-art models. The second section has results for Llama-8B, and the third one for Llama 70B. Underlined scores denote best in section per metric while bold scores denote best overall. AlignScore-R calculates AlignScore against source reviews, while AlignScore-M is computed against reference meta-reviews.

faithfulness (see AlignScore). Our aspect-aware decomposition is consistently superior to more naive decompositions and prompting methods in terms of aspect coverage across domains and model backbones. We also observe that using Llama-70B as a backbone gives our approach a boost across metrics which is not surprising as larger models tend to have better generation and instruction-following capabilities. Interestingly, the fine-tuned model (FT-Llama 8B) trails behind our modular system when using a backbone LLM of the same scale (Aspect-aware decomposition-Llama 8B), both in terms of aspect coverage and opinion faithfulness. Overall, our results suggest that prompt decomposition is useful in opinion summarization and intermediate reasoning steps based on task and domain-specific knowledge lead to meta-reviews of higher quality.

Llama-70B performs well at identifying and summarizing aspects. In addition to evaluating the generated meta-reviews, we conduct evaluations on the intermediate outputs of our modules. We only report results on the scientific domain reusing the ground truth annotations[12] provided in Li et al. (2024). For *Aspect Identification*, we calculate word-level Recall, Precision, and F_1 between model-extracted text fragments and human-annotated text fragments following Li et al. (2024). The scores shown in Table 5 denote how accurately our approach extracts opinionated text from source reviews. We find that Llama-3.1-70B is the best model for this module, even better than GPT-4o (in terms of F_1). Moreover, Figure 2 shows that Llama-3.1-70B also performs well on individual review aspects, especially frequent ones including *Novelty*, *Soundness* and *Clarity*. For *Opinion Consolidation*, Table 6 shows that Llama-3.1-70B performs better than other models at generating aspect-specific meta-reviews. Taken together, the evaluations on intermediate outputs explain Llama-3.1-70B's superior performance at the end task.

[12] https://github.com/oaimli/MetaReviewingLogic

Models	Recall↑	Precision↑	F_1 ↑
GPT-4o	**0.82**	0.27	0.40
Llama-3.1-8B	0.80	0.25	0.38
Llama-3.1-70B	0.74	**0.34**	**0.46**

Table 5: Evaluation of text fragments extracted by *Aspect Identification* against human annotations.

Models	AlignScore-S↑	Rouge↑
GPT-4o	0.86	18.40
Llama-3.1-8B	0.82	**18.24**
Llama-3.1-70B	**0.87**	16.93

Table 6: Evaluation of aspect-specific meta-reviews, i.e., intermediate outputs of *Opinion Consolidation*.

***Opinion Consolidation* is the most important module.** We further examine the contributions of individual modules to meta-review generation. Specifically, we perform two ablations: (1) remove *Aspect Identification* and directly generate aspect-specific meta-reviews based on original reviews and (2) remove *Opinion Consolidation* and directly generate final meta-reviews based on text fragments from *Aspect Identification*. We use Llama-3.1-70B as our backbone LLM because of its superior performance in previous experiments. As we have ground truth text fragments for scientific reviews (Li et al., 2024), we include another experiment in this domain where we replace the output of *Aspect Identification* with human-annotated text fragments. According to Table 7, both *Aspect Identification* and *Opinion Consolidation* are crucial to generating more faithful meta-reviews and with higher aspect coverage, however *Opinion Consolidation* appears to be the most critical as its removal decreases performance across domains (exception: coverage for research articles). We also see an interesting observation where model-extracted text fragments are on par with human-selected ones but more helpful to generating faithful meta-reviews.

Humans prefer meta-reviews generated by our modular system to gold-standard references. We conduct a human evaluation to verify that our approach generates meta-reviews that reflect the review aspects of the input and are overall coherent and faithful. We recruited crowdworkers through Prolific[13], selected to be L1 English speakers from the US or UK, and compensated above the UK living wage at 12GBP/hr. We ask crowdworkers to read a set of source reviews followed by two generated meta-reviews and select which meta-review is best (allowing for ties) along two dimensions, as well as an overall preference:

- **Coverage** — Which meta-review covers more review aspects in the source reviews?
- **Faithfulness** — Which meta-review has a higher percentage of opinions supported by the source reviews?
- **Overall** — Which meta-review do you think is better overall?

We randomly select ten entities for each dataset (SPACE, AmaSum, and PeerSum) and construct six pairwise combinations between our approach (Aspect-aware decomposition with Llama-3.1-70B) and the systems shown in Table 8, including human-written reference meta-reviews. For AmaSum and SPACE, we only present crowdworkers with 20% of the reviews for each entity, to maintain a reasonable workload (reviews are sampled randomly). We elicit three annotations for each pairwise combination of system outputs, leading to

[13] https://www.prolific.com/

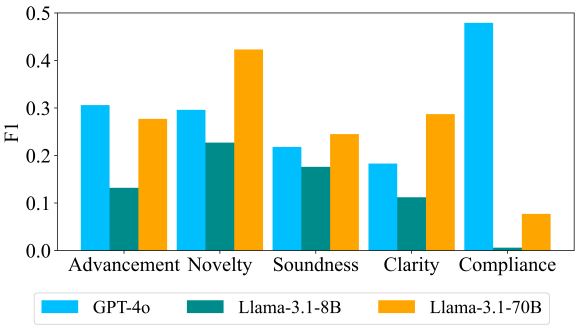

Figure 2: Evaluation of text fragments extracted for individual review aspects by *Aspect Identification*.

Domain	Modules	Coverage↑	AlignScore-S↑
Hotels	AI+OC+MS	**0.99**	0.80
	OC+MS	**0.99**	**0.83**
	AI+MS	0.55	0.62
Shoes	AI+OC+MS	**0.83**	**0.74**
	OC+MS	0.69	0.72
	AI+MS	0.61	0.69
Research Articles	AI+OC+MS	0.97	**0.79**
	OC+MS	**0.98**	0.78
	AI+MS	0.97	0.75
	AI†+OC+MS	0.97	0.69

Table 7: Ablations quantifying the contribution of different modules on three domains (hotels, shoes, research articles). AI: *Aspect Identification*, OC: *Opinion Consolidation*, MS: *Meta-Review Synthesis*, AI†: text fragments selected by humans. Results shown for *Aspect-aware decomposition-Llama 70B*.

Model	Cover↑	Faith↑	Overall↑
Research Articles			
Sentiment CoT-GPT-4o	0%	0%	0%
Human-written reference	80%	80%	80%
Automatic decomposition-Llama 70B	90%	90%	90%
Chunk-wise decomposition-Llama 70B	70%	90%	90%
Naive aspect-aware prompting-Llama 70B	0%	0%	10%
Aspect-aware decomposition-GPT-4o	10%	50%	50%
Sports Shoes			
HIRO-abs	90%	90%	90%
Human-written reference	90%	90%	90%
Automatic decomposition-Llama 70B	100%	90%	100%
Chunk-wise decomposition-Llama 70B	80%	80%	70%
Naive aspect-aware prompting-Llama 70B	20%	20%	40%
Aspect-aware decomposition-GPT-4o	10%	20%	30%
Hotels			
HIRO-abs	80%	100%	100%
Human-written reference	30%	70%	100%
Automatic decomposition-Llama 70B	90%	100%	100%
Chunk-wise decomposition-Llama 70B	50%	60%	80%
Naive aspect-aware prompting-Llama 70B	100%	100%	100%
Aspect-aware decomposition-GPT-4o	0%	0%	10%

Table 8: Proportion of times (%) crowdworkers preferred our model (*Aspect-aware decomposition-Llama 70B*) against depicted systems. We highlight in red comparisons where our model is chosen as better more than 50% of the time (higher is better). For example, '90%' means that crowdworkers prefer our system on 9 out of 10 entities. We take a majority vote to determine a single system preference.

a total of 1,260 ratings. Annotators have reasonable agreement, with average values of Krippendorff's α being 0.335 on shoes, 0.622 on hotels, and 0.463 on research articles. More details on experimental design and the full instructions are in Appendix G.

Table 8 shows the proportion of times (%) crowdworkers prefer our approach against a comparison system. We find that human judgments are broadly consistent with automatic evaluation. Crowdworkers prefer our system to human references on two (shoes and research articles) out of three domains. We consistently win against automatic and chunk-wise decompositions (with Llama 70B), but lose against our own decompositions with GPT-4o.

Aspect-aware decomposition allows humans to create better summaries faster. We also evaluate the intermediate outputs produced by our modules. In particular, we examine whether the specific module decomposition adopted by our system is useful for real-world meta-review writing. We ask annotators to write meta-reviews for hotel reviews in three conditions: (1) they are not given any intermediate reasoning steps; (2) they are given reasoning steps produced by automatic knowledge-agnostic decomposition from *Automatic decomposition-Llama 70B*; and (3) they are provided with the intermediate outputs of our modules with *Aspect-aware decomposition-Llama 70B* as reasoning steps. We record the time it takes crowdworkers to finish the writing.

Present Reasoning Steps	Time↓	Preferred↑
No reasoning steps	10.9	20%
Automatic decomposition	10.3	20%
Aspect-aware decomposition (ours)	9.3	40%

Table 9: Average time (in minutes) humans take to write scientific meta-reviews and the proportion of times participants prefer meta-reviews when present with different intermediate reasoning steps (in exhausted pair-wise comparison).

We randomly select ten entities and obtain three meta-reviews for each (according to the three conditions described above). We recruited five annotators, however, each annotator writes a meta-review for each entity once to avoid memorization. Based on the time reported in Table 9, we find that providing intermediate outputs of our aspect-aware decomposition accelerates participants' writing compared with the other two conditions and it reduces the time of writing a meta-review by 14.7% (on average). More details about how we present different reasoning steps to annotators and annotation instructions are provided in Appendix H. We also ask another set of annotators to assess the meta-reviews written above, by presenting pair-wise comparisons (following the instructions of human annotation presented in the previous section). We find that participants prefer meta-reviews written based on the outputs of our modules twice as much compared to the other two settings (Krippendorff's α is 0.542).

6 Conclusion

We propose modular decomposition for opinion summarization based on review aspects. Our decomposition is evidence-based (the output of each module can be traced back to its input), enabling greater transparency and ease of inspection. Extensive experiments demonstrate that our modular framework outperforms state-of-the-art methods and other strong baselines in multiple domains. Human evaluations reveal that our approach not only produces higher-quality meta-reviews but also generates more useful intermediate outputs to assist humans in composing meta-reviews. While our work focuses on opinion summarization, the concept of aspect-aware decomposition holds promise for other complex language generation tasks.

Limitations

Our work, while promising, has some limitations. Firstly, all three experimental datasets used in our study are in English, limiting the evaluation to a single language. Secondly, the prompts for our modular approach could be further optimized, as we did not focus extensively on prompt optimization. Finally, our approach does not explicitly address the potential generation of biased or harmful content, even though our goal is to ensure that the generated meta-reviews remain grounded in the original reviews.

Ethics Statement

Our work primarily focuses on enhancing the capabilities of AI systems to assist humans, rather than aiming to replace them. As demonstrated in our experiments, the intermediate outputs generated by our approach can help humans produce higher-quality meta-reviews with greater efficiency.

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.

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.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

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.

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

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

Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. 2023. From key points to key point hierarchy: Structured and expressive opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–928, Toronto, Canada. 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.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

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.

Tom Hosking, Hao Tang, and Mirella Lapata. 2024. Hierarchical indexing for retrieval-augmented opinion summarization. *Transactions of the Association for Computational Linguistics*, 12:1533–1555.

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

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Walter Chang, and Fei Liu. 2020. A cascade approach to neural abstractive summarization with content selection and fusion. 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 529–535, Suzhou, China. Association for Computational Linguistics.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. 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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

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.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI Blog*.

Aviv Slobodkin, Ori Shapira, Ran Levy, and Ido Dagan. 2024. Multi-review fusion-in-context. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3003–3021, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Prompts for Aspect-aware Decomposition

In this section we provide the prompt templates used to decompose opinion summarization into the modules of *Aspect Identification*, *Opinion Consolidation*, and *Meta-review Synthesis*. Domain-specific prompts are provided in Sections B–D.

Aspect Identification

You are good at understanding documents with {domain} review opinions.
Below is a {domain} review for an academic manuscript, please extract fragments that are related to {the-review-aspect} of the {the entity}.
Definition of {the review aspect}:{the definition of the review aspect}
Example input review:
{the example input review}
Example format of extracted fragments in different lines:
{the example output}
Target input review:
{input-document}
Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 3: The few-shot prompt template for the *Aspect Identification* module; text fragments are extracted for each (domain) aspect. Please note that for research articles we use few-shot prompting to enable the model follow the output format while for sports shoes and hotels zero-shot prompting (with just removing the demonstration example) could get reasonable performances.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.
Example input review fragments:
{the example text fragments}
Example summary of the input fragments:
{the example aspect-specific meta-review of the input fragments}
Target input fragments:
{input-fragments}
The final summary of these target input text fragments (just output the answer without any other content):

Figure 4: The few-shot prompt template for the *Opinion Consolidation* module; it outputs summaries for individual review aspects. Please note that for research articles we use few-shot prompting to get better performance while for sports shoes and hotels zero-shot prompting (with just removing the demonstration example) could get reasonable performances.

Meta-Review Synthesis

You are good at understanding documents with {domain} review opinions.
Below are comments on different review aspects for {the entity}, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.
Comments on different aspects:
{meta-reviews of individual review aspects}
The meta-review is (directly output the answer without any other content):

Figure 5: The prompt template for the *Meta-Review Synthesis* module based on aspect-specific meta-reviews from the *Opinion Consolidation* module. As zero-shot prompting gives us reasonable performances on all the three datasets, we used the same zero-shot prompt template for the module.

B Prompts for Scientific Reviews of Research Articles

Prompts for *Aspect Identification* are given in Tables 6–10 for the aspects *Advancement*, *Clarity*, *Compliance*, *Soundness*, and *Novelty*. The prompt for *Opinion Consolidation* is in Table 11 and all aspects share the same prompt for this module. The prompt for *Meta-Review Synthesis* is in Table 12.

Aspect Identification: Advancement

You are good at understanding documents with scientific review opinions.
Below is a scientific review for an academic manuscript, please extract text fragments that are related to Advancement of the research work.

Definition of Advancement:

Importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

Example input review:

This paper theoretically studied one of the fundamental issue in CycleGAN (recently gained much attention for image-to-image translation). The authors analyze the space of exact and approximated solutions under automorphisms. Reviewers mostly agree with theoretical value of the paper. Some concerns on practical values are also raised, e.g., limited or no-surprising experimental results. In overall, I think this is a boarderline paper. But, I am a bit toward acceptance as the theoretical contribution is solid, and potentially beneficial to many future works on unpaired image-to-image translation.

Example output fragments in different lines:

Some concerns on practical values are also raised, e.g., limited or no-surprising experimental results.

Reviewers mostly agree with theoretical value of the paper.

But, I am a bit toward acceptance as the theoretical contribution is solid, and potentially beneficial to many future works on unpaired image-to-image translation.

Target input review:

{{input_document}}
Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 6: The prompt of *Aspect Identification* for the aspect *Advancement*.

C Prompts for Business Reviews of Hotels

Prompts for *Aspect Identification* on hotels are shown in Tables 13–18 for the aspects *Building*, *Cleanliness*, *Food*, *Location*, *Rooms*, and *Service*. The prompt for *Opinon Consolidation* for any review aspect is in Table 19. The prompt for *Meta-Review Synthesis* is present in Table 20.

D Prompts for Product Reviews of Sports Shoes

Prompts for *Aspect Identification* are given in Tables 21–30 for the aspects *Breathability*, *Comfort*, *Cushioning*, *Durability*, *Flexibility*, *Misc*, *Size and Fit*, *Stability*, *Traction*, and *Weight*. The prompt for *Opinion Consolidation* for any aspect is in Table 31. The prompt for *Meta-Review Synthesis* is in Table 32.

E Implementation Details of Comparison Models

In this section we provide implementation details for the various comparison models used in our experiments.

- For HIRO-abs (Hosking et al., 2024), we obtain generations for AmaSum and SPACE from https://github.com/tomhosking/hiro. There are three outputs for each entity and we use the first one as the generation of HIRO-abs.
- For fine-tuning Llama-3.1-8B, we trained the model with Transformers from Huggingface on the three datasets for 5 epochs on four NVIDIA A100 80G GPUs, with `max-predict-length=512`, `bf16=True`, `batch-size=1`, `optim=adafactor`, `learning-rate=1e-6`, `warmup-rate=0.2`, `label-smoothing-factor=0.1`, `lr-scheduler-type=cosine`, `fsdp='full_shard auto_wrap offload'`.

Aspect Identification: Clarity

You are good at understanding documents with scientific review opinions.
Below is a scientific review for an academic manuscript, please extract fragments that are related to Clarity of the research work.

Definition of Clarity:

The readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

Example input review:

The paper is about a software library that allows for relatively easy simulation of molecular dynamics. The library is based on JAX and draws heavily from its benefits.

To be honest, this is a difficult paper to evaluate for everyone involved in this discussion. The reason for this is that it is an unconventional paper (software) whose target application centered around molecular dynamics. While the package seems to be useful for this purpose (and some ML-related purposes), the paper does not expose which of the benefits come from JAX and which ones the authors added in JAX MD. It looks like that most of the benefits are built-in benefits in JAX. Furthermore, I am missing a detailed analysis of computation speed (the authors do mention this in the discussion below and in a sentence in the paper, but this insufficient). Currently, it seems that the package is relatively slow compared to existing alternatives.

Here are some recommendations:
1. It would be good if the authors focused more on ML-related problems in the paper, because this would also make sure that the package is not considered a specialized package that overfits to molecular dynamics.
2. Please work out the contribution/delta of JAX MD compared to JAX.
3. Provide a thorough analysis of the computation speed.
4. Make a better case, why JAX MD should be the go-to method for practitioners.

Overall, I recommend rejection of this paper. A potential re-submission venue could be JMLR, which has an explicit software track.

Example output fragments in different lines:

While the package seems to be useful for this purpose (and some ML-related purposes), the paper does not expose which of the benefits come from JAX and which ones the authors added in JAX MD.

Make a better case, why JAX MD should be the go-to method for practitioners.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 7: The prompt of *Aspect Identification* for the aspect of *Clarity*.

- For *naive aspect-aware prompting*, we only incorporate aspect descriptions into the prompt. As an example, we show the prompt for scientific reviews in Figure 33.
- For *Automatic decomposition* (Khot et al., 2023), the prompting approach cannot be directly transferred to opinion summarization. Based on the idea of automatic decomposition, we implement automatic knowledge-agnostic decomposition on our experimental datasets. The idea is to first generate intermediate reasoning steps and then follow those steps in sequence to generate the final meta-review. We provide example prompts for scientific reviews in Figure 34 and 35.
- For *chunk-wise decomposition* (Khot et al., 2023), we first generate small meta-reviews for each chunk, and then combine all chunk-specific meta-reviews with another prompt to generate the global meta-review. Example prompts for scientific reviews are shown in Figures36 and 37.

F Implementation Details for Automatic Evaluation

Implementation details of G-Eval (Liu et al., 2023) are presented in Figures 38, 39, and 40 for the three domains, respectively. We use gpt-4o-2024-05-13 as the backbone LLM of G-Eval.

G Details of Human Evaluation on Quality of Generated Meta-Reviews

We conduct human evaluation based on pair-wise comparisons to verify the quality of our generated meta-reviews (in terms of aspect coverage and opinion faithfulness). We recruited crowdworkers through

Aspect Identification: Compliance

You are good at understanding documents with scientific review opinions.
Below is a scientific review for an academic manuscript, please extract fragments that are related to Compliance of the research work.

Definition of Compliance:

Whether the manuscript fits the venue, and all ethical and publication requirements are met.

Example input review:

"The paper proposes a method to identify and correct regions on the data manifold in which a trained classifier fails. The *identification* phase is based on clustering classification failure regions in a GAN latent space and the *correction* phase is based on fine-tuning the classifier with additional synthetic samples from the GAN. The proposed method is strongly based on Zhao et al 2018 (Generating Natural Adversarial Examples), a method to generate on-manifold black-box adversarial examples using a GAN. The authors of the current paper describe some differences of their identification step from Zhao et al (end of section 3.2.1), but in my opinion they are minor. The main contribution of the current paper over Zhao et al seems to be clustering the adversarial examples (using GMM) and using them to fine-tune the classifier. This, in my opinion, is potentially an interesting idea, however, the authors do not show sufficient evidence of its success. Specifically, the authors claim to "achieve near perfect failure scenario accuracy with minimal change in test set accuracy", but they do not provide any details (e.g. table of accuracy values on the train, test and adversarial sets before and after the fine-tuning). I would also expect to see an ablation study comparing the proposed method to simply including the adversarial examples found using Zhao et al (w/o GMM fitting and sampling) as additional training example - a standard adversarial defense approach (see e.g. [1]).Perhaps more importantly, the objective of the proposed method is not, in my opinion, clear. The title and abstract describe the goal as "debugging" a classifier and correcting fail regions, however the described method seems like a defense against on-manifold adversarial attack. If the method, as claimed, helps debugging and correcting the classifier, I would expect to see an improved accuracy on the (natural) unseen test set - not just on the synthetically generated adversarial examples. The quality and clarity of the writing can be improved as well. A lot of space is allocated to describing well-known methods (e.g. VAE, GMM), however, critical information about the experimental results are missing. I'm also not sure all the formally defined algorithms and equations actually help in the understanding (e.g. algorithm 1, equation 2). Some of the mathematical notations are not standard. Minor comment: The norm in definition 3.1 is a regular vector norm (l2?) and not a matrix norm. To summarize: pros: - interesting idea (clustering on-manifold failures, labeling them and then using them to improve the classifier)cons:- contribution over Zhao et al not well established- insufficient and inaccurate experimental results- general quality of writing - not sure actual work and experiments match the stated objective - significance *Update:* Following the authors' response, I upgraded my rating, but I still think there are critical issues with the paper. The most problematic point, in my opinion, is the only-marginal improvement on the test data, indicating that the suggested training method only improves the specific "failure scenarios", making it is similar to adversarial training methods used to gain adversarial robustness. However, the abstract and introduction indicates that the paper helps in debugging in fixing failures in general, which, I think should have been evident in improved test accuracy.[1] Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy."ICML 2019

Example output fragments in different lines:

Some of the mathematical notations are not standard.

Target input meta-review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 8: The prompt of *Aspect Identification* for the aspect of *Compliance*.

Aspect Identification: Soundness

You are good at understanding documents with scientific review opinions.
Below is a scientific meta-review for an academic manuscript, please extract fragments that are related to Soundness of the research work.

Definition of Soundness: There are usually two types of soundness: (1) Empirical: how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted. (2) Theoretical: whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology (e.g., mathematical approach) and the analysis is correct.

Example input meta-review:

The paper proposes to use the mirror descent algorithm for the binary network. It is easy to read. However, novelty over ProxQuant is somehow limited. The theoretical analysis is weak, in that there is no analysis on the convergence and neither how to choose the projection for mirror mapping construction. Experimental results can also be made more convincing, by adding comparisons with bigger datasets, STOA networks, and ablation study to demonstrate why mirror descent is better than proximal gradient descent in this application.

Example output fragments in different lines:

The theoretical analysis is weak, in that there is no analysis on the convergence and neither how to choose the projection for mirror mapping construction.

Experimental results can also be made more convincing, by adding comparisons with bigger datasets, STOA networks, and ablation study to demonstrate why mirror descent is better than proximal gradient descent in this application.

Target input meta-review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 9: The prompt of *Aspect Identification* for the aspect of *Soundness*.

Aspect Identification: Novelty

You are good at understanding documents with scientific review opinions.
Below is a scientific meta-review for an academic manuscript, please extract fragments that are related to Novelty of the research work.

Definition of Novelty:

How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

Example input meta-review:

The manuscript describes a method for identifying and correcting classifier performance when labels are assigned incorrectly. The identification is based on clustering classification failure regions in a VAE latent space and the correction phase is based on fine-tuning the classifier with additional synthetic samples from the VAE.

Reviewers agreed that the manuscript is not ready for publication. The main issue is that the suggested training method is similar to adversarial training methods used to gain adversarial robustness. The method does not help in debugging and fixing failures in general.

Example output fragments in different lines:

Reviewers appreciated the novelty, introducing a new simpler routing mechanism, and achieving good performance on real world datasets.

In particular, removing the squash function and experimenting with concurrent routing was highlighted as significant progress.

Alongside with them, I acknowledge the novelty of using layer norm and parallel execution, and recommend accept.

Target input meta-review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 10: The prompt of *Aspect Identification* for the aspect of *Novelty*.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.
Example input review fragments:

"1) **Evaluating different explanation techniques:**",
"We thus believe that our results do *not* violate the surmise made in the shared reference, but rather support it.",
"We believe this makes our findings generalizable.",
"Although, the paper brings out the importance of analogies as explanations (which further motivates our work)",
"The proposed technique is flexible as it can provide two forms of explanations: feature and analogy-based.",
"Moreover, explanations in the form of analogies are intuitive for human users.",
"We feel that analogous examples do not need to share common words, content, or sentence structure. What is important is that they *point to latent factors* that may be responsible for the model's output.",
"**Purpose of analogies:**",
...
"The authors solved this problem by the use of a learned local distance matrix, in which interaction effects are clearly shown."

Example summary of the input fragments:

The proposed approach to explain similarity prediction is a relatively less explored area, which makes the problem addressed and the proposed method unique.

Example input review fragments:
"The paper is technically sound, and the claims are carefully developed and well supported.",
"The manuscript is well structured and very clearly written, with helpful introductions to the methodological ingredients that it builds upon.",
"The paper could be further improved with some reflection on the limitations of the approach.",
"I am not certain how large a contribution it will have to the field of Bayesian inference in general.",
...
"I'll use the rest of the section for high-level comments.",
"- In its current form, the paper convinces me that SHF decreases runtime and increases performance for datasets with low complexity."

Example summary of the input fragments:

Based on these, I recommend acceptance for this paper. All reviewers agree that the paper proposes an interesting approach to Bayesian inference incorporating coresets with Hamiltonian flows.

Target input review fragments:

{{review_fragments}}

The final summary of these target input text fragments (just output the answer without any other content):

Figure 11: The prompt of *Opinion Consolidation* for any aspect of scientific reviews.

Meta-review Synthesis

You are good at understanding documents with scientific review opinions.
Below are comments on different review aspects for an academic manuscript, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.

Comments on different aspects:

{{metas_generated}}

The meta-review is (directly output the answer without any other content):

Figure 12: The prompt of *Meta-Review Synthesis* for research articles.

Aspect Identification: Building

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Building of the hotel.

Definition of Building:
Analysis of how well the hotel was constructed, its design, functionality, and how these factors contribute to the success and satisfaction of its guests.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 13: The prompt of *Aspect Identification* for the aspect of *Building*.

Aspect Identification: Cleanliness

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Cleanliness of the hotel.

Definition of Cleanliness:
Evaluation of how well the hotel maintains a clean, sanitary, and comfortable environment for its guests, impacting their overall experience and satisfaction.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 14: The prompt of *Aspect Identification* for the aspect of *Cleanliness*.

Aspect Identification: Food

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Food of the hotel.

Definition of Food:
Evaluation of the dining experience including the quality and variety of the food, ultimately affecting guest satisfaction and the hotel's reputation.
Target input review:
{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 15: The prompt of *Aspect Identification* with the aspect of *Food*.

Prolific[14] with compensation above the UK living wage at £12 per working hour.

For product reviews of sports shoes, we randomly select ten entities from the test data of AmaSum. Based on generated meta-reviews, for each entity we construct six pairs of comparisons between our modular approach with Llama-3.1-70B as a backbone and comparison baselines. There are originally about 400 source reviews in each entity and it is hard for humans to review all of them. To balance annotator workload, we present annotators with 20% reviews and randomly select reviews for three times to ensure experimental consistency. Therefore, there are 18 pairs of comparisons for each entity. Each

[14] www.prolific.com

Aspect Identification: Location

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Location of the hotel.

Definition of Location:
Analysis of how the hotel's location influences the guest experience, considering factors like convenience, safety, proximity to attractions, and the overall environment.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 16: The prompt of *Aspect Identification* for the aspect of *Location*.

Aspect Identification: Rooms

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Rooms of the hotel.

Definition of Rooms:
Assessment of how well the room meets the guest's needs and expectations in terms of comfort, cleanliness, amenities, and overall experience.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 17: The prompt of *Aspect Identification* for the review aspect of *Rooms*.

Aspect Identification: Service

You are good at understanding documents with hotel review opinions.
Below is a business review for a hotel, please extract fragments that are related to Service of the hotel.

Definition of Service:
Assessment of how well the hotel staff and management meet the needs of their guests, impacting their comfort, convenience, and overall experience.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 18: The prompt of *Aspect Identification* with the aspect of *Service*.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.

Target input review fragments:

{{review_fragments}}

The final summary of these target input text fragments (just produce the answer without any other content):

Figure 19: The prompt of *Opinion Consolidation* for any individual review aspect for hotels.

Meta-Review Synthesis

You are good at understanding documents with hotel review opinions.
Below are business reviews in different aspects for a hotel, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.

Comments on different aspects:

{{metas_generated}}
The meta-review is (directly output the answer without any other content):

Figure 20: The prompt of *Meta-Review Synthesis* for hotels.

Aspect Identification: Breathability

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Breathability of shoes.

Definition of Breathability:
Evaluation about breathability of the shoes.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 21: The prompt of *Aspect Identification* for the aspect of *Breathability*.

Aspect Identification: Comfort

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Comfort of shoes.

Definition of Comfort:
Evaluation about comfort of the shoes, such as tongue padding, heel tab, and removable insole.

Target input review:
{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 22: The prompt of *Aspect Identification* with the aspect of *Comfort*.

Aspect Identification: Cushioning

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Cushioning of shoes.

Definition of Cushioning:
Evaluation about cushioning of the shoes, such as heel stack and forefoot stack.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 23: The prompt of *Aspect Identification* for the review aspect of *Cushioning*.

Aspect Identification: Breathability

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Durability of shoes.

Definition of Durability:
Evaluation about durability of the shoes, such as outsole hardness and thickness.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 24: The prompt of *Aspect Identification* with the aspect of *Durability*.

pair is rated by three different annotators and we obtain 540 annotations for the dataset.

We recruited 27 annotators from Prolific with L1 English from the US or UK, with a minimum approval rate of 100% in more than 100 studies. In addition to the attention check question for each annotation instance, we also included quality control instances, asking participants to distinguish human-written reference meta-reviews from random meta-reviews (taken from other entities). Each annotator worked on 20 annotation instances for the main study and another 4 quality control instances. Raters were asked five questions about review aspects and opinion faithfulness. Our annotation instructions and interface are shown in Figure 41, Figure 42, and Figure 43. After filtering out annotators failing more than one

Aspect Identification: Flexibility

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Flexibility of shoes.

Definition of Flexibility:
Evaluation about flexibility of the shoes, such as stiffness, stiffness in the cold, and difference in stiffness in the cold.

Target input review:

{{input_document}}
Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 25: The prompt of *Aspect Identification* with the review aspect of *Flexibility*.

Aspect Identification: Misc

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Misc of shoes.

Definition of Misc:
Evaluation about reflective elements of the shoes.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 26: The prompt of Aspect Identification with the review aspect of Misc.

Aspect Identification: Size and Fit

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Size and Fit of shoes.

Definition of Size and Fit:
Evaluation about size and fit of the shoes, such as internal length, toebox width at the widest part, and gusset type.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 27: The prompt of *Aspect Identification* for the aspect of *Size and Fit*.

Aspect Identification: Stability

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Stability of shoes.

Definition of Stability:
Evaluation about stability of the shoes, such as torsional rigidity, heel counter stiffness, midsole width in the forefoot and midsole width in the heel.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 28: The prompt of *Aspect Identification* for the aspect of *Stability*.

quality control annotation pair, the annotators have reasonable agreement and the average Krippendorff's α of 0.335.

We follow the same setting for the evaluation of meta-reviews for hotels. There are also 540 annotations, and we obtain 27 annotators from Prolific. The annotation instructions and experimental interface are

Aspect Identification: Traction

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Traction of shoes.

Definition of Traction:
Evaluation about traction of the shoes, such as lug depth.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 29: The prompt of *Aspect Identification* for the review aspect of *Traction*.

Aspect Identification: Weight

You are good at understanding documents with sports shoes review opinions.
Below is a product review for a pair of shoes, please extract fragments that are related to Weight of shoes.

Definition of Weight:
Evaluation about weight of the shoes.

Target input review:

{{input_document}}

Final extracted fragments (follow the format above in different lines and if no resulted fragments just output "No related fragments"):

Figure 30: The prompt of *Aspect Identification* for the review aspect of *Weight*.

Opinion Consolidation

You are good at writing summaries for opinionated texts. You are given some opinionated text fragments, please write a concise summary for them.
Target input review fragments:

{{review_fragments}}

The final summary of these target input text fragments (just produce the answer without any other content):

Figure 31: The prompt of *Opinion Consolidation* for any individual review aspect for sports shoes.

Meta-Review Synthesis

You are good at understanding documents with sports shoes review opinions.
Below are product reviews in different aspects for a pair of shoes, please write a concise and natural meta-review which summaries the provided comments and covers all mentioned review aspects.

Comments on different aspects:

{{metas_generated}}

The meta-review is (directly output the answer without any other content):

Figure 32: The prompt of *Meta-Review Synthesis* for the product reviews of sports shoes.

shown in Figure 44, Figure 45, and Figure 46. After filtering out annotators who failed on more than one quality control instances, the average Krippendorff's α is 0.622.

For scientific reviews of research articles, we randomly select ten entities from the test data of PeerSum. There are also six pairs of comparisons between our modular approach with Llama-3.1-70B as a backbone and comparison baselines. As there are only about 15 reviews on average, we show annotators all reviews. Therefore, there are 6 pairs of comparisons for each entity. Each pair gets annotated by three different annotators and we have 180 annotations for the dataset. We elicited 9 annotators from Prolific with

Naive Aspect-Aware Prompt

Please write a summary for the reviews on a scientific article, focused on the review aspects below.

Review aspects:

(1) Advancement: importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

(2) Clarity: the readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

(3) Compliance: whether the manuscript fits the venue, and all ethical and publication requirements are met.

(4) Soundness: there are usually two types of soundness, empirical (how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted.) and theoretical (whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology, e.g., mathematical approach and the analysis is correct.)

(5) Novelty: how original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

Reviews on a scientific article:
{{source_documents}}

The output summary:

Figure 33: The prompt with aspects in scientific reviews of research articles for *naive aspect-aware prompting*.

Automatic Decomposition Prompt

You are requested to write the steps. Please output the final answer with only the steps in different lines, no other useless content.

Please give me sequential steps to write a summary specific for the following reviews on an academic paper.
Reviews on a paper: {source_text}
The steps to write a summary in different lines:

Figure 34: The prompt for automatic decomposition to generate intermediate reasoning steps to write the meta-review for scientific reviews.

Prompt to Follow Reasoning Steps from Automatic Decomposition

You are requested to follow the instruction and only generate the requested output.

{output_from_last_step}
Please follow the instruction below and give your output.
{current_step}
The output:

Figure 35: The prompt to follow automatically predicted steps by *automatic decomposition* to generate the final meta-review.

required L1 English from the US or UK, and a minimum approval rate of 100% in more than 100 studies. We also required that they are pursuing a PhD in computer science or engineering. In addition to the attention check question for each annotation instance, we also included quality control instances, same as before. Therefore, each annotator worked on 20 pairs of comparisons for the main study and another 4 quality control instances. In each annotation, participants are asked 5 questions about review aspects and

Chunk Summarization Prompt

You are requested to do summarization. Please output the final answer with only the summary, no other useless content.

Please write a summary for the following review on an academic paper.
The review: {the_text_chunk}
The output summary:

Figure 36: The prompt of *chunk-wise decomposition* to summarize individual chunks of texts for scientific reviews of research articles.

Summary Aggregation Prompt

You are requested to do summarization. Please output the final answer with only the summary, no other useless content.

Please write a summary for the following texts.
The texts to be summarized:
{the_concatenation_of_small_meta_reviews_of_chunks}
The output summary:

Figure 37: The Prompt for aggregating chunk-specific meta-reviews into the global meta-review.

G-Eval for Sports Shoes

Here are several review documents that contain opinions from different people about a pair of shoes, along with a candidate summary of these reviews.

You are required to evaluate how accurately the given summary reflects the overall opinions for review aspects expressed in the original reviews.

Please read all opinions in the summary and calculate the percentage of faithful opinions that are clearly supported by the source review documents.

Review documents:

{{source_documents}}

The candidate summary:

{{generation_summary}}

The percentage of faithful opinions (only output a decimal like 0.12, no other content):

Figure 38: The G-Eval prompt for evaluating meta-reviews for sports shoes.

opinion faithfulness. The annotation instructions and interface are shown in Figure 47, Figure 48, and Figure 49. After filtering out annotators failing more than one quality control instances, the annotators, the average Krippendorff's α is 0.463.

H Details of Human Evaluation on Usefulness of Intermediate Outputs

To record the time that humans spend to write meta-reviews with different reasoning steps, we conduct the experiments also with Prolific and present annotators interfaces with instructions in Figure 50, Figure 51 and Figure 52.We recruited five crowdworkers through Prolific[15] with compensation above the UK living wage at £12 per working hour. These annotators are required to be experienced in L1 English from the US or UK, with a minimum approval rate of 100% in more than 100 studies. Annotators are required to focus

[15]www.prolific.com

G-Eval for Research Articles

Here are several review documents that contain opinions from different people about a scientific paper, along with a candidate summary of these reviews.

You are required to evaluate how accurately the given summary reflects the overall opinions for review aspects expressed in the original reviews.

Please read all opinions in the summary and calculate the percentage of faithful opinions that are clearly supported by the source review documents.

Review documents:

{{source_documents}}

The candidate summary:

{{generation_summary}}

The percentage of faithful opinions (only output a decimal like 0.12, no other content):

Figure 39: The G-Eval prompt for evaluating meta-reviews on research articles.

G-Eval for Hotels

Here are several review documents that contain opinions from different people about a hotel, along with a candidate summary of these reviews.

You are required to evaluate how accurately the given summary reflects the overall opinions for review aspects expressed in the original reviews.

Please read all opinions in the summary and calculate the percentage of faithful opinions that are clearly supported by the source review documents.

Review documents:

{{source_documents}}

The candidate summary:

{{generation_summary}}

The percentage of faithful opinions (only output a decimal like 0.12, no other content):

Figure 40: The G-Eval prompt for evaluating meta-reviews on hotels.

on the annotation task and finish the writing task in a continuous period of time. The study is conducted on ten entities and there are three meta-reviews for each (according to the three conditions described in Section 5). To avoid memorization, each annotator must write a meta-review for each entity only once. We find that all our annotators passed our attention check question present in our instructions Figure 52. We calculate the average time that the participants take for the ten instances in each condition from the five annotators.

To compare the quality of written meta-reviews in the three different conditions, we run another human evaluation in the same setting as the one to compare model-generated meta-reviews in Section 5. This was also based on pair-wise comparison and there were 30 pairs of comparison. We recruited three annotators and each pair of comparison was annotated for three times. The agreement among the three annotators is high (Krippendorff's α is 0.542).

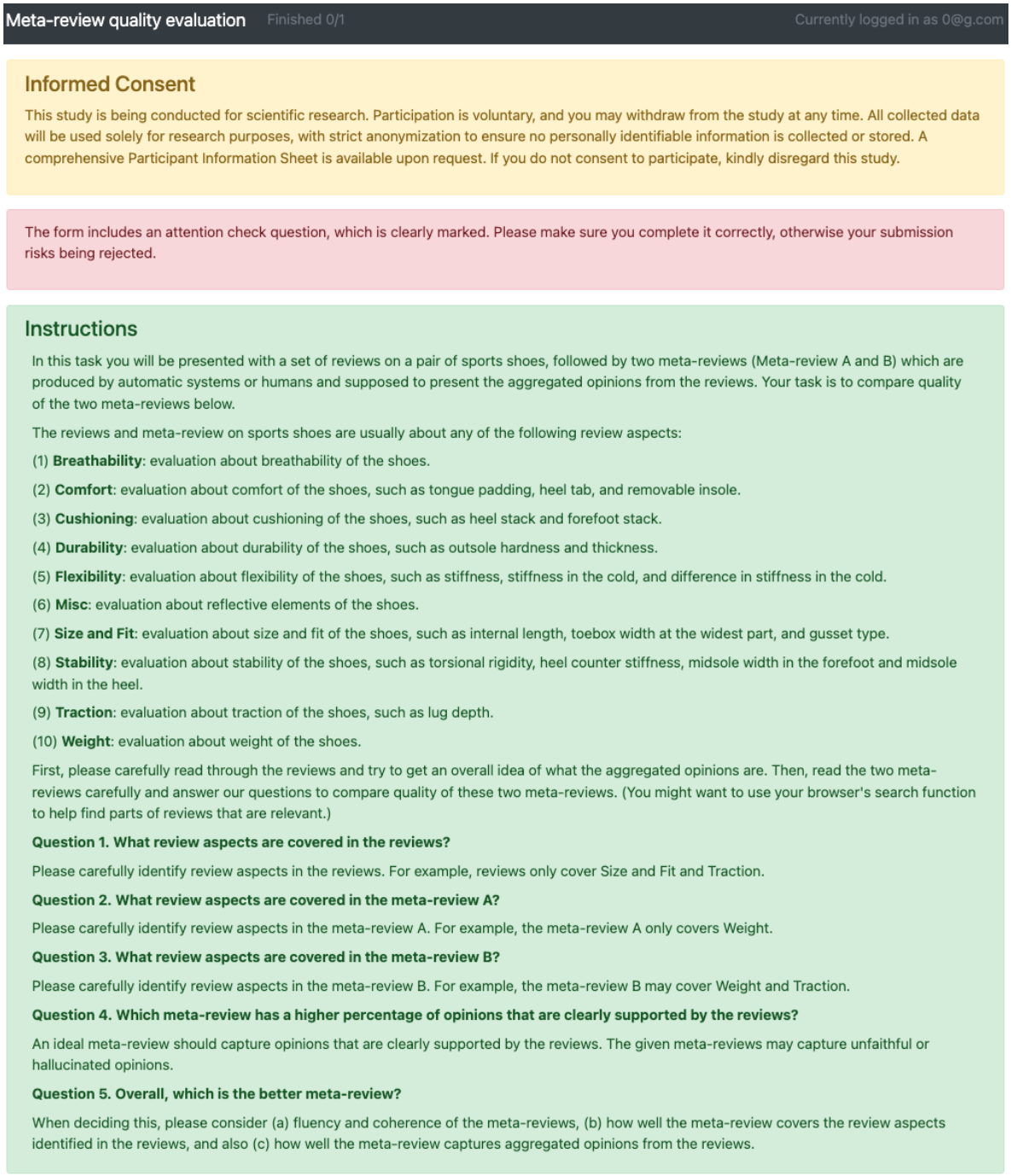

Figure 41: Experimental instructions and interface for human evaluation study on sports shoes reviews (part 1).

Figure 42: Experimental instructions and interface for human evaluation study on sports shoes (part 2).

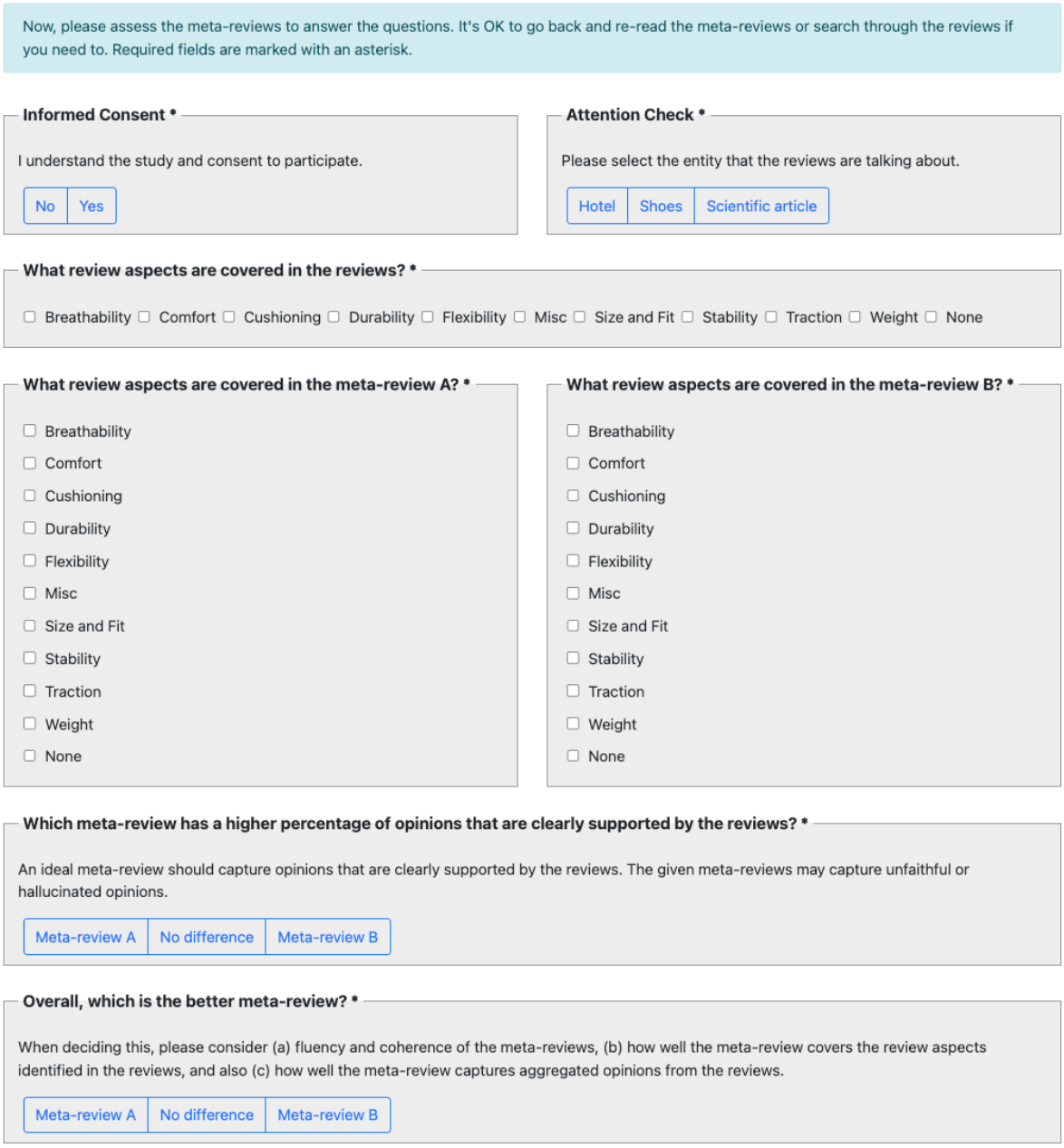

Figure 43: Experimental instructions and interface for human evaluation study on sports shoes (part 3).

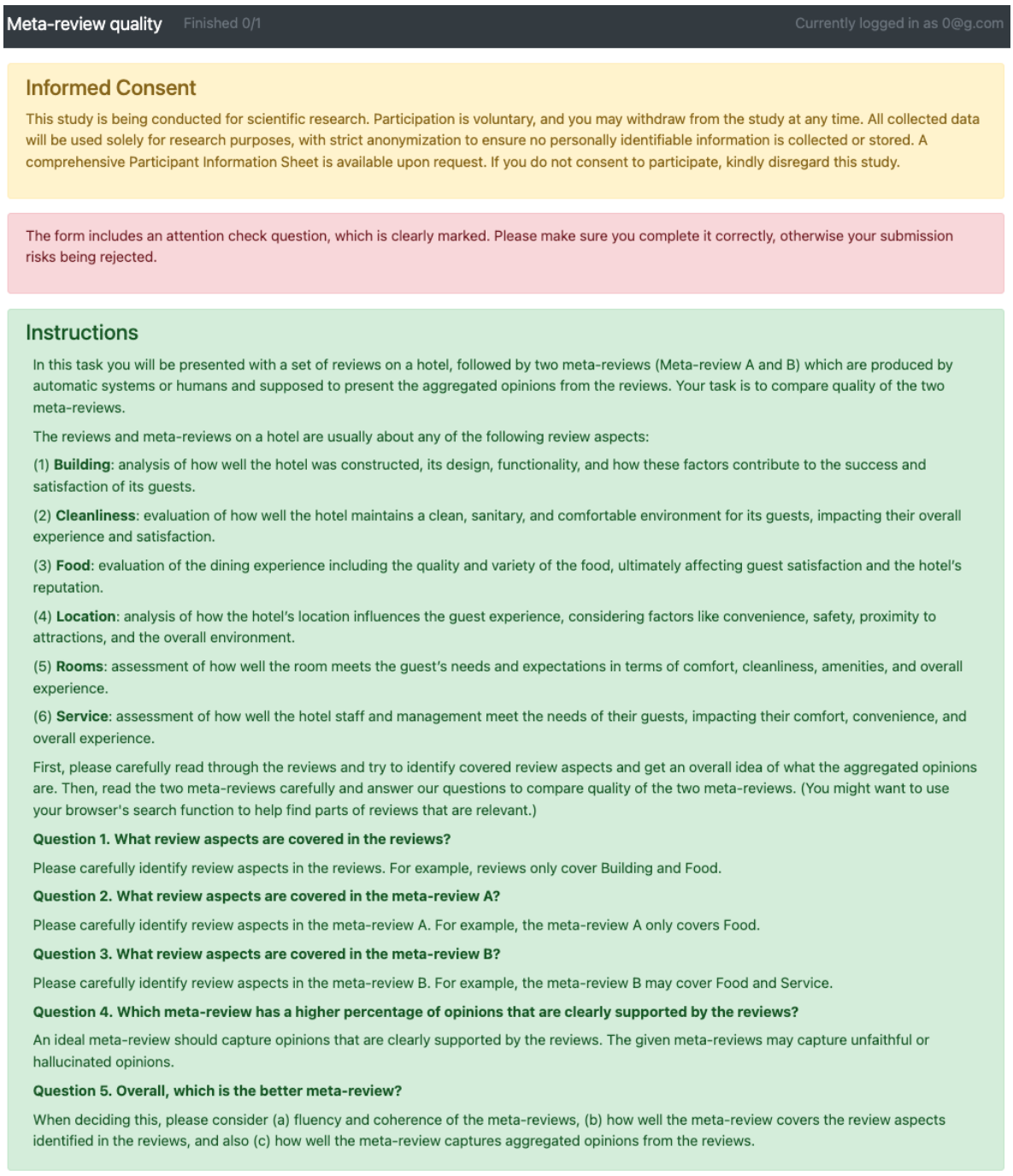

Figure 44: Experimental instructions and interface for human evaluation study on hotels (part 1).

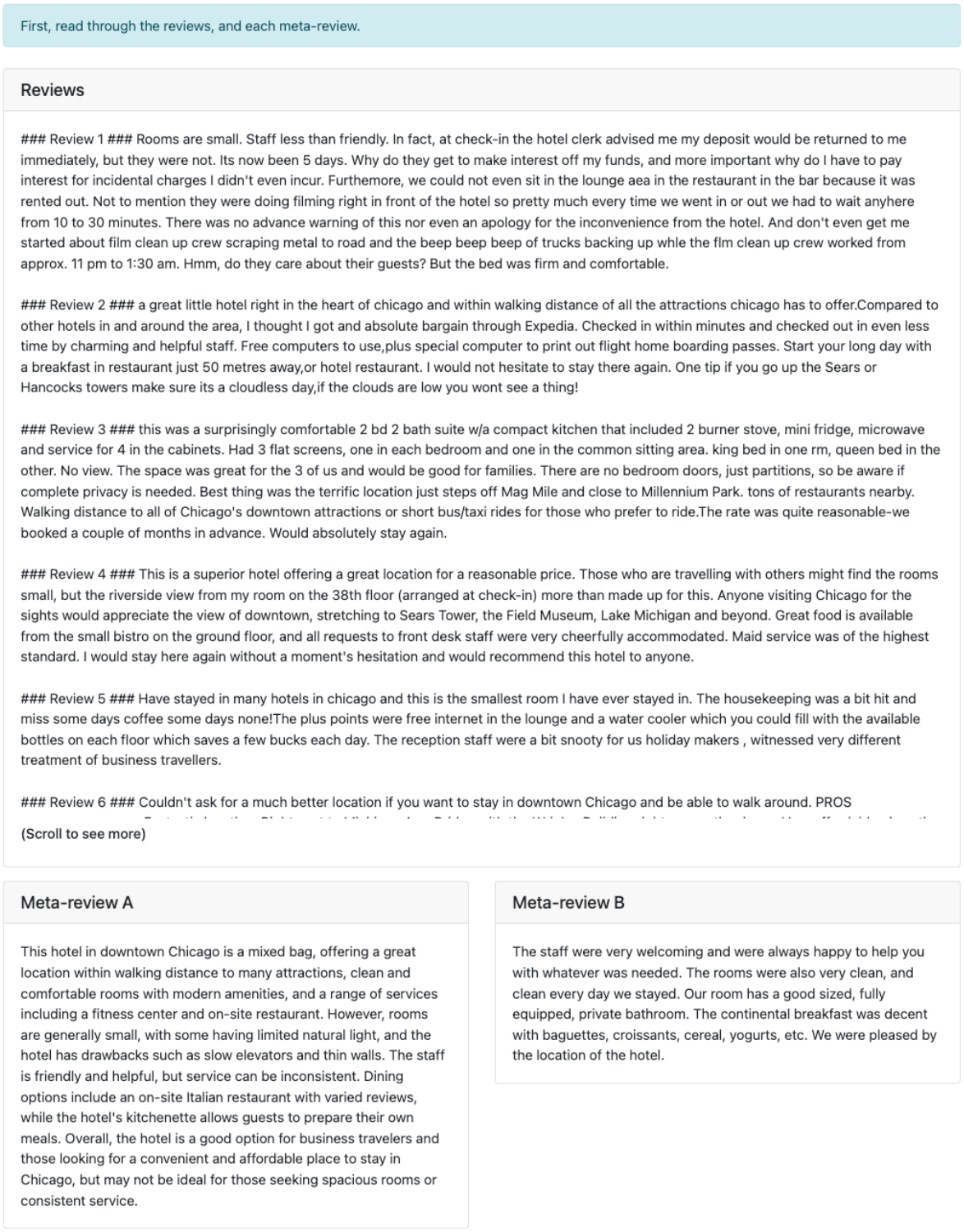

Figure 45: Experimental instructions and interface for human evaluation study on hotels (part 2).

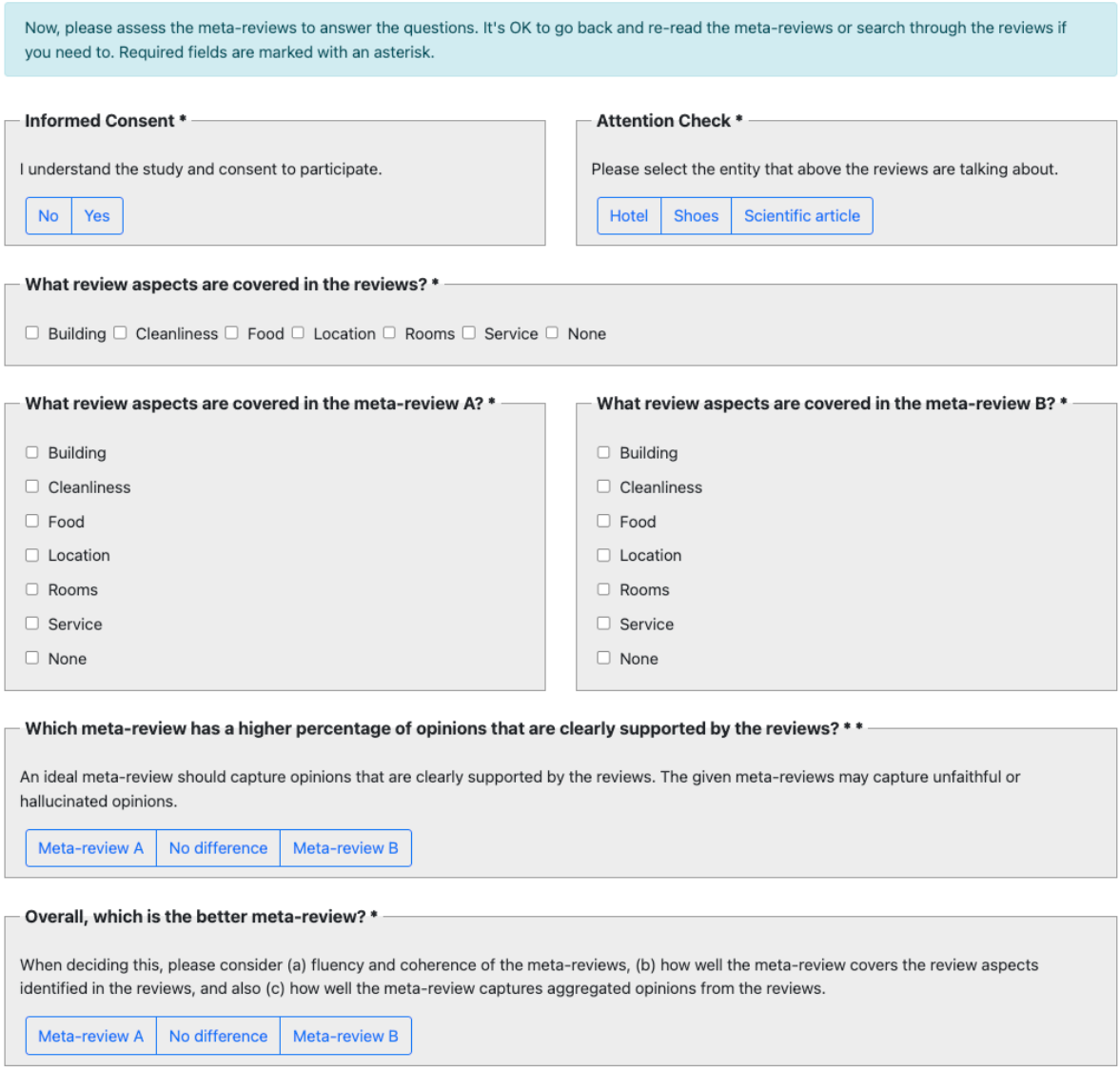

Figure 46: Experimental instructions and interface for human evaluation study on hotels (part 3).

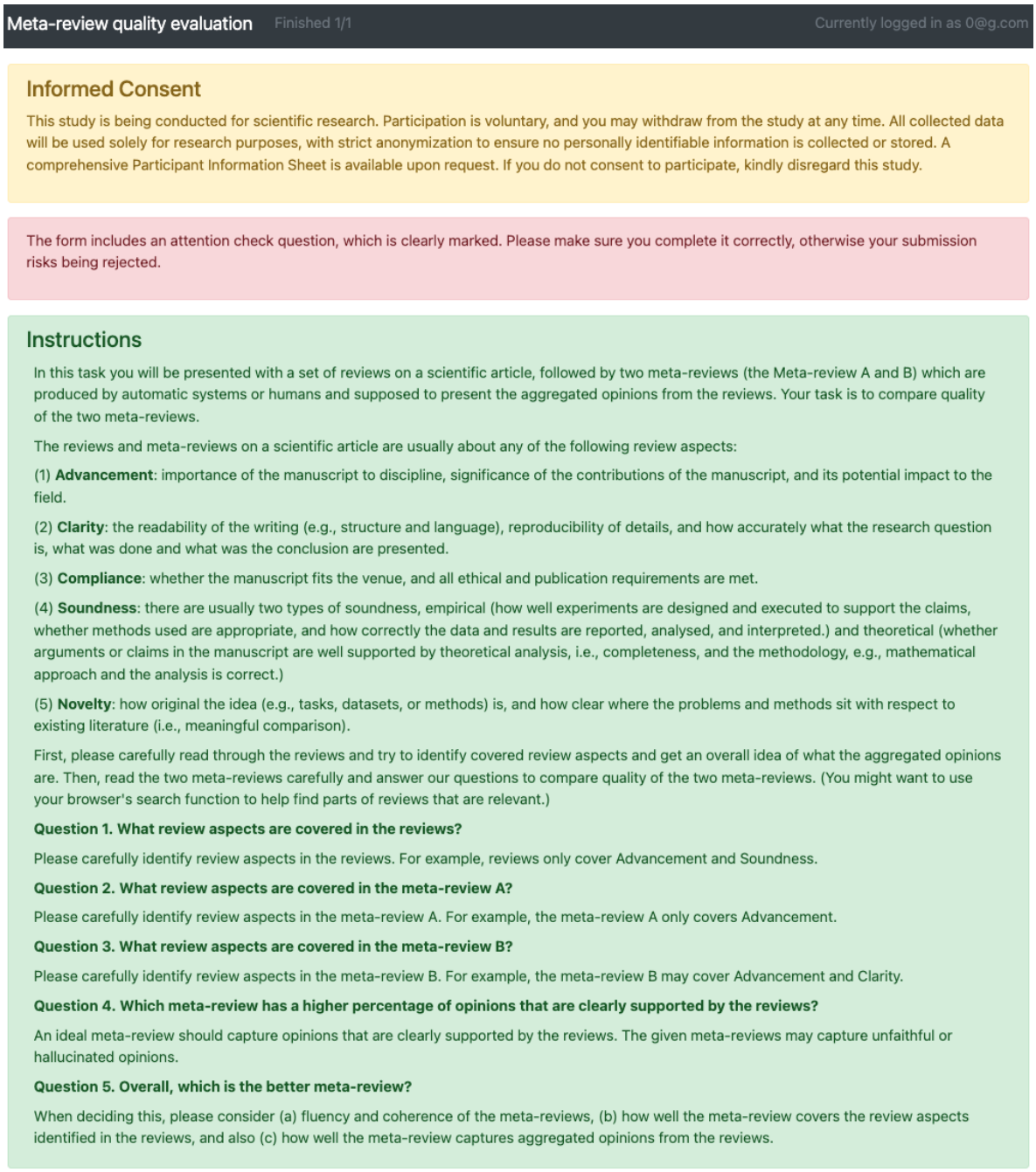

Figure 47: Experimental instructions and interface for human evaluation study on article reviews (part 1).

First, read through the reviews, and two meta-reviews.

Reviews

The Paper Abstract

We evaluate the information that can unintentionally leak into the low dimensional output of a neural network, by reconstructing an input image from a 40- or 32-element feature vector that intends to only describe abstract attributes of a facial portrait. The reconstruction uses blackbox-access to the image encoder which generates the feature vector. Other than previous work, we leverage recent knowledge about image generation and facial similarity, implementing a method that outperforms the current state-of-the-art. Our strategy uses a pretrained StyleGAN and a new loss function that compares the perceptual similarity of portraits by mapping them into the latent space of a FaceNet embedding. Additionally, we present a new technique that fuses the output of an ensemble, to deliberately generate specific aspects of the recreated image.

The Review 1

This paper studies the unintentional information leakage that can happen in deep encoder networks that extract latent representations with abstract attributes from face images. The paper proposes a method that is capable to reconstruct an input face image from a feature vector representation using only black box access to the image encoder. The method is based on the StyleGAN formulation, which is extended with an additional loss that compares the perceptual similarity of portraits by mapping them into the latent space of a FaceNet embedding. The purpose of this paper is to raise awareness about the relevant security issues of existing deep learning systems for face analysis. + This paper deals with an interesting and important problem that has attracted limited attention from the computer vision community. It is particularly important for reasons related to security and preservation of privacy.

+ The proposed pipeline is intuitive and sound, building upon the formulation of the StyleGAN model. - The technical novelty of the proposed method is relatively limited. It only describes a small extension of the loss function of the StyleGAN model. It is mostly interesting as an application of the GAN-based formulations, but I think that it lacks sufficient contributions for a paper accepted in ICLR. Other venues might be more appropriate for such paper.

- The experimental evaluation is highly inadequate. The only quantitative evaluation is the one presented in Table 1. However, this corresponds to an internal evaluation of the proposed method, without any comparison with other SOTA methods. Closely related methods like (Yang et al., 2019) and (Zhao et al. 2021) should have been included in the quantitative comparisons. In addition, a perceptual user study should have been included in the experiments, in order to quantify the performance of the proposed method and other compared methods, in terms of whether the reconstructed faces are perceived by humans to have the same identity as the original real faces.

- The paper has also inadequacies in terms of discussing and citing prior art. First, Some closely-related works, like (Razzhigaev et al. 2020) are only presented in Table 2 of the Appendix. However, such works should have been presented in the main paper, with discussion about their similarities and differences from the proposed method. Furthermore, the paper has not cited some closely-related works like the following:

(Scroll to see more)

Meta-review A

This manuscript proposes a novel method for reconstructing a target face image from a low-dimensional feature vector, addressing an important problem related to security and privacy preservation in the computer vision community. While the approach is interesting and leverages recent knowledge in image generation and facial similarity, outperforming the current state-of-the-art, the paper has several significant inadequacies. The experimental evaluation is inadequate, lacking comparison with state-of-the-art methods and clear conclusions, which raises questions about the validity of the findings. Additionally, the discussion of prior art is insufficient, and the structure and content of the paper are not suitable for this venue. The authors need to provide more justification and ablation studies for their approach to strengthen the manuscript.

Meta-review B

The paper proposes a learning method (specifically a deep equilibrium learning approach) for 'regularization by denoising', a plug-and-play method for solving inverse problems.

After the rebuttal, all reviewers support acceptance of the paper. The reviewers find the paper to be well written, the problem to be interesting, and the claims to be well supported (reviewer Hjnn), both empirically (reviewer uDGc) and through theory. Reviewer A7f5 finds the work particularly exciting since both memory and training time are reduced, without sacrificing image quality.

Based on my own reading and the unanimous support of the reviewers, I recommend acceptance of the paper. A nice contribution!

Figure 48: Experimental instructions and interface for human evaluation study on article reviews (part 2).

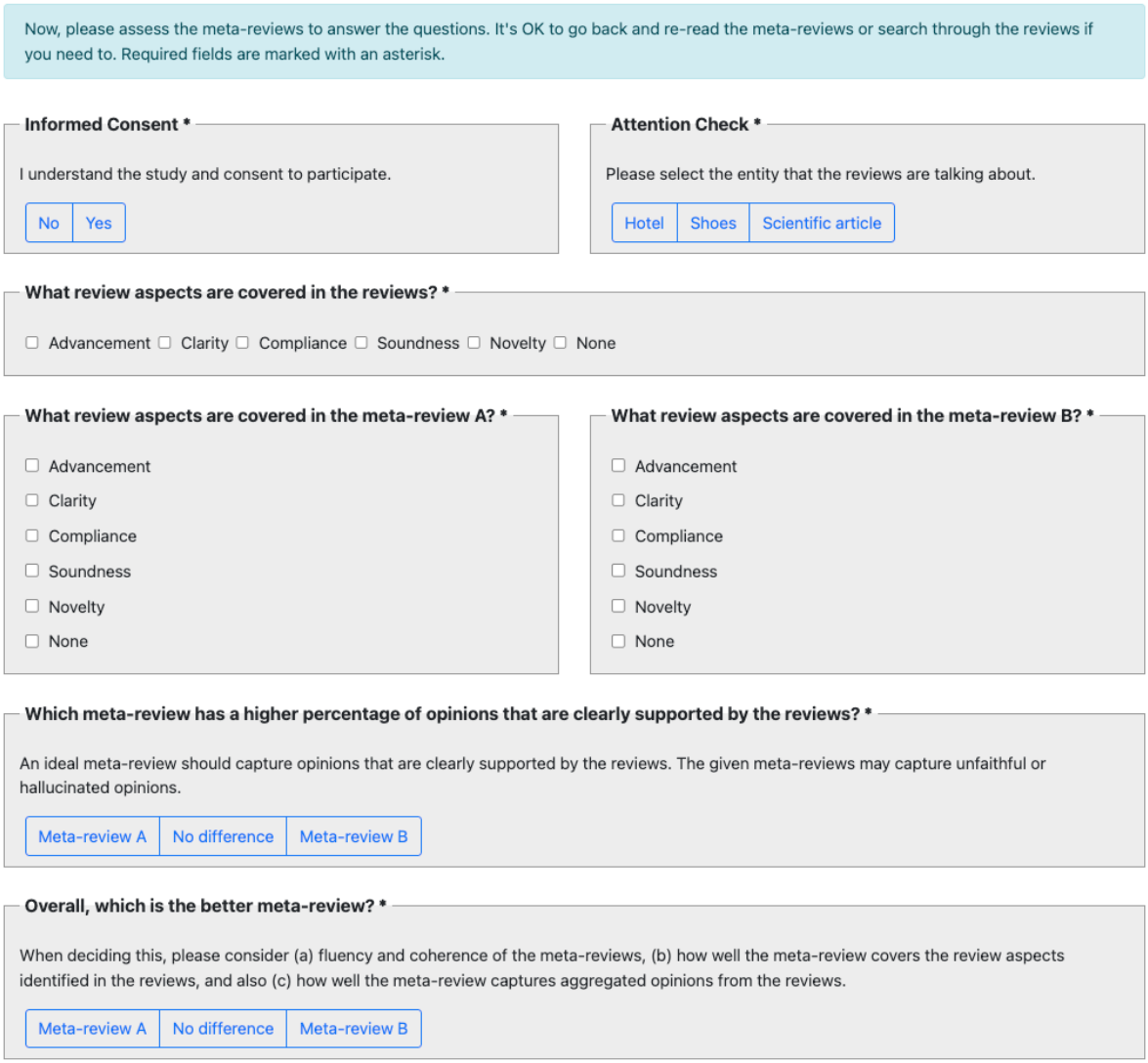

Figure 49: Experimental instructions and interface for human evaluation study on article reviews (part 3).

Figure 50: Interface for annotators to write meta-reviews based on different intermediate outputs (part 1).

for a small price break for spa treatments to Roman Spa Resort guests. As it is folks that come in off the street pay the same rates for spa treatments as resort guests. My suggestion is make it more enticing for a resort guest to do a spa treatment especially ones that are staying multiple nights. All in all, I have no complaints. The Resort is very clean, the grounds and landscaping are fantastic, with a Spanish/Mission style motif. There is adequate parking and everything in town is within walking distance.

We spent 4 days at the Roman Spa on our honeymoon and had a most wonderful time. Be aware - this is not one of the big hotel chains so no fancy high tech facilities, no wi-Fi and no restaurant What you do get is VERY comfortable accommodation, kitchenette - excellent little supermarket around the corner so you can eat in (Healthier & cheaper) without restrictions on menus etc. The staff were helpful and friendly and the spa is for being thoroughly spoiled! And the location is close to everything

(Scroll to see more)

Intermediate Steps

You could write the meta-review based on aggregation of aspect-focused meta-reviews that we provide below if you find them useful. You will see an aspect-focused meta-review and corresponding text fragments extracted from the reviews.

############## Review Aspect ##############
Building
############## Aspect-Focused Meta-Review ##############
The resort offers clean, comfortable, and well-maintained rooms and facilities, including multiple heated pools, a patio area, and kitchenettes. Although it lacks high-tech amenities like Wi-Fi and a restaurant, its Spanish/Mission style grounds and landscaping are well-kept and beautiful. The location is convenient, with everything in town within walking distance. However, some rooms may need decor updates and the beds can be hard.
############## Extracted Text Fragments ##############
1. The facilities are kept in perfect condition.
2. These is a beautiful patio area with eating tables and Weber BBQ [bricketts] as well as a gas cooking facility.
3. They have three heated pools, one large outdoor, medium indoor and a hot tub type thing.
4. The rooms are clean, modern, and are kept up.
5. The Resort is very clean, the grounds and landscaping are fantastic, with a Spanish/Mission style motif.
6. There is adequate parking and everything in town is within walking distance.
7. kitchenette - excellent little supermarket around the corner so you can eat in
8. VERY comfortable accommodation,
9. Be aware - this is not one of the big hotel chains so no fancy high tech facilities, no wi-Fi and no restaurant What you do get is VERY comfortable accommodation, kitchenette - excellent little supermarket around the corner so you can eat in (Healthier & cheaper) without restrictions on menus etc.
10. the three therapy pools also beautifully kept with grounds and flowers
11. The rooms are emaculate and well appointed
12. The grounds are simply amazing!
13. Pots of tulips and daffodils in full bloom; other plantings well cared for; pathways clean and swept.
14. Our room was clean and comfortable
15. While the rooms could use a style update, ours was clean and had a small but nice bathroom.
16. Our room had a kitchenette which was convenient, but since the spa is located only steps away from a variety of restaurants (high, medium and low end), we just used it for the refrigerator and early morning coffee.
17. The bed was HARD!
18. It needs an update, new decor, the whole 9.
19. The room was very clean.
20. The room was also dark, even with the curtains open, so we had to have lights on all the time.

(Scroll to see more)

Please answer the following questions and write a meta-review based on your understanding of the reviews. It's OK to go back and re-read the reviews or search through them if you need to. Required fields are marked with an asterisk.

Please (1) make sure you correctly complete the attention check question which is clearly marked, (2) do not use any AI tools for writing, (3) do not directly use extracted sentences as the meta-review, and (4) finish the writing in a continuous period of time, otherwise your submission risks being rejected.

Figure 51: Interface for annotators to write meta-reviews based on different intermediate outputs (part 2).

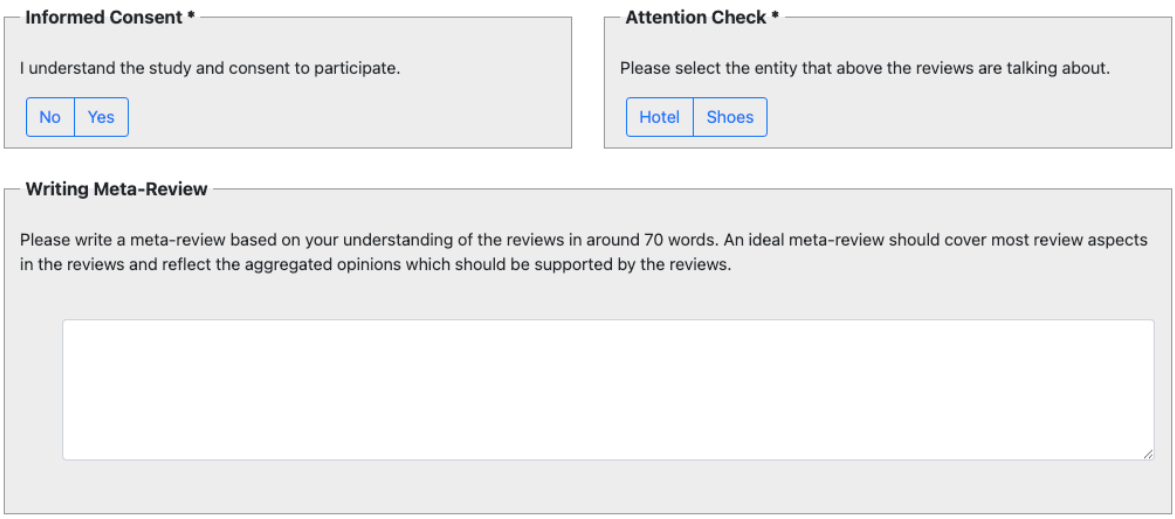

Figure 52: Interface for annotators to write meta-reviews based on different intermediate outputs (part 3).