Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks

Ruiyang Zhou[1] Lu Chen[1,2,✉] Kai Yu[1,2,✉]
[1] X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence, SJTU AI Institute
Shanghai Jiao Tong University, Shanghai, China
[2] Suzhou Laboratory, Suzhou, China
{ellenruiyang,chenlusz,kai.yu}@sjtu.edu.cn

Abstract

The use of large language models (LLM), especially ChatGPT, to help with research has come into practice. Researchers use it for timely advice and hope to obtain in-depth feedback. However, can LLM be a qualified and reliable reviewer? Although there already exist several review-related datasets, few works have carefully and thoroughly inspected model's capability as a reviewer, especially the correctness of generated reviews. In this paper, we first evaluate GPT-3.5 and GPT-4 (the current top-performing LLM) on 2 types of tasks under different settings: the score prediction task and the review generation task. In addition, we propose a dataset containing 196 review-revision multiple-choice questions (RR-MCQ) with detailed labels from the review-rebuttal forum in ICLR-2023. By asking questions from technical details to the overall presentation and quality, our RR-MCQ data provides a more complete model ability assessment. The results show that LLM is generally helpful, but great caution is needed as it always makes mistakes. Although it can give passable decisions (> 60% accuracy) on single options, completely correct answers are still rare (about 20%); models are still weak on long paper processing, zero-shot scoring, and giving critical feedback like human reviewers.

Keywords: automatic peer review, large language model, multiple choice question answering

1. Introduction

Utilizing large language models for scientific paper review recently attracts researcher's interest. The continuously growing amount of new paper publications, together with the increasing specialization within various research fields makes it a challenge to obtain timely and in-depth feedback. At the same time, LLM demonstrates strong ability in reading comprehension, knowledge integration, and even logical reasoning (OpenAI, 2023). Thus arises naturally this question: **can LLM be a qualified and reliable automatic reviewer?**

In fact, even before the release of (truly) large language models, there already exist datasets and methods targeting review-related tasks. For example, finetuning pretrained models to predict paper decision and review scores (Li et al., 2020), or using language models to generate review texts (Yuan et al., 2022). Lately, new datasets for review generation and edit generation also appear (D'Arcy et al., 2023), but there is still no detailed assessment of model's reviewing ability.

In this paper, we first examine the reviewing ability of GPT-3.5 and GPT-4 from two perspectives: review aspect score prediction and review generation. The two types of tasks evaluate both the ability to discover flaws in research papers and rectify them, from the granularity level of abstract scoring to detailed commenting. We take great caution during the evaluation process due to the innate difficulty of evaluating freely generated texts: besides classical automatic metrics, new metrics and manual evaluations are also implemented.

We then design a "qualification exam" for fine-grained analysis: we construct 196 review-revision-related multiple-choice questions. On previous datasets like review generation ones, detailed analyses are only possible when manually examining the generated text, resulting in huge time costs and subjective conclusions. Even with manual analysis, the correctness of generated reviews is still difficult to measure. In contrast, our RR-MCQ dataset with well-defined categorization labels enables comprehensive and satisfactory assessments. The questions are inferred from real discussion forums of 55 reviews from 14 papers in ICLR-2023, investigating both criticizing and correcting abilities. Due to the high cost of designing high-quality questions, we limit the total number of questions to about 200 (196 to be specific).[1]

We come to the following conclusions:

- LLM has the potential to give meaningful scores and decide on individual statements.
- However, they are NOT easy to use in practice: seldom fully correct, not critical enough,

✉ Lu Chen and Kai Yu are corresponding authors.

[1] Our RR-MCQ data is available at https://huggingface.co/datasets/zhouruiyang/RR-MCQ

LREC-COLING 2024, pages 9340–9351
20-25 May, 2024.

lack technical details, and struggle with long context.

- Automatic similarity metrics do not align with the true review generation quality; the assessment of model reliability is needed.

2. Related Work

2.1. Paper-reviewing Related Task

Generation task Automatic review generation is the most direct task in using models as automatic reviewers. Datasets like PeerRead (Kang et al., 2018), ASAP (Yuan et al., 2022), ReviewRobot (Wang et al., 2020), MOPRD (Lin et al., 2023), and NLPEER (Dycke et al., 2022) all contain scientific papers (mostly in the domain of computer science) and their corresponding peer reviews. However, since it is difficult to directly generate review texts and evaluate them, various types of annotations have been proposed. The most common label is the sentence type, classified based on the sentence's sentiment polarity, review aspect, or the text aspect that it comments on, for example COMPARE (Singh et al., 2021), ReAct (Choudhary et al., 2021), AMSR (Fromm et al., 2021), COMPARE (Singh et al., 2021), Peer-Review-Analyze (Ghosal et al., 2022), and AMPERE (Hua et al., 2019). Still, there is no commonly recognized best annotation style and evaluation metric.

Other generation tasks related to reviewing ability also appear, like meta-review generation and edit (revision) generation. The meta-review generation task is more similar to the summarization task, but it summarizes multiple peer reviews in the scientific field. MreD (Shen et al., 2021) is an exemplified meta-review generation task dataset with sentence intent annotations. The revision generation task is more complicated, requiring the ability to comprehend the comment as well as take actions; examples are Revise and Resubmit (Kuznetsov et al., 2022), ArxivEdits (Jiang et al., 2022) and ARIES (D'Arcy et al., 2023).

Classification task Besides directly generating texts, more specific tasks with clear-cut answers are actually more investigated. Paper decision prediction and aspect score prediction are the two most researched tasks before the appearance of large language models, like in PeerRead (Kang et al., 2018). Other reviewing-related tasks include argument extraction in RR (Cheng et al., 2020) and DISAPERE (Kennard et al., 2021), sentence classification (on datasets mentioned above for review generation with annotations).

In this paper, we first evaluate models on both types of existing tasks: the generation task and the classification task. Specifically, we choose to run GPT-3.5 and GPT-4 on the review generation task and aspect score prediction task; the result of GPT-4 on edit generation is already presented in ARIES (D'Arcy et al., 2023). We then present our RR-MCQ data that inspects all aspects mentioned above.

2.2. Large Language Models for Reviewing

Recently, Liu and Shah (2023) inspects GPT-4's ability by constructing a small-sized artificial test dataset. They first create 13 short papers and then design test examples based on these brand-new papers to avoid the data leakage problem. They find that GPT-4 can accomplish the list-checking task, but makes frequent mistakes on error-identifying and paper-ranking tasks. Their detailed analysis is only based on a limited number of manually designed questions; in contrast, our RR-MCQ dataset has more test questions whose distribution basically aligns with reality.

ARIES (D'Arcy et al., 2023) proposes a dataset for comment-edit pairing and edit generation task, but find that even GPT-4 aligns badly the comment and the edit, and that the GPT-4 generated revisions have low coherence and insufficient technical details. However, they do not measure the correctness of generated revisions as it is extremely difficult. Our work turns the free-generation task into a multiple-choice question-answering task, making the measurement of correctness easy and automatic. It is like a qualification test for LLM before being an automatic reviewer.

Robertson (2023) tests the usefulness of GPT-4 generated reviews by questioning 10 real users. Very recently, Liang et al. (2023) assess on a larger scale the GPT-4 generated reviews: they tag the comment overlap (hit rate and several other overlap coefficients) and survey the user satisfaction to measure the review quality. They find that the reviews have satisfactory overlap and consistency with human references, but can be non-generic and emphasize different aspects. Comment overlap is an important indicator, but our RR-MCQ data offers more evaluation perspectivesof the model's ability and reliability.

3. Task 1: Aspect Score Prediction

3.1. PeerRead Dataset

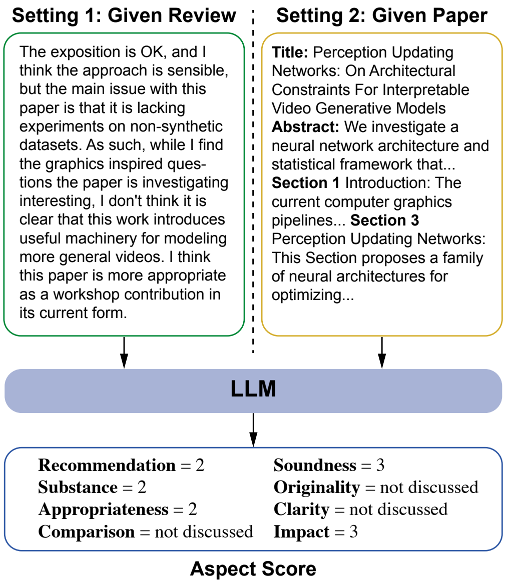
For the task of aspect score prediction, we use the ICLR-2017 subset of the PeerRead dataset (Kang et al., 2018). This subset contains 1.3k manually annotated aspect scores (ranging from 1 to 5 inclusive) for 427 official reviews from ICLR-2017 conference. The manual annotations ensure the feasibility and consistency of the aspect score prediction task: aspects that are not discussed in the review have a special *not discussed* score label. See Figure 1 for a concrete example.

		1. accuracy ↑	2. |diff| ↓	3. Pearson ↑	4. Spearman ↑	5. Kendall's tau ↑
baseline	1. most frequent score	***0.404***	0.966	0.333	0.340	0.297
given review	2. zero-shot	**0.353**	***0.856***	0.548	0.553	0.475
	3. few-shot	0.306	1.132	***0.651***	***0.659***	***0.580***
	4. MCQ style	0.336	1.025	0.558	0.565	0.492
given paper	5. abstract	0.237	0.992	0.228	0.233	0.195
	6. whole paper (GPT-3.5-16k)	0.138	2.132	0.131	0.131	0.109
	7. selected sections	0.251	**0.886**	**0.258**	**0.265**	**0.222**
	8. abstract & sections	**0.330**	0.923	0.248	0.249	0.209

Table 1: Average results of the aspect score prediction task from GPT-3.5 and GPT-3.5-16k on PeerRead dataset. ↑ means the higher the metric value, the better the performance. The best result under each setting is bolded, and the best score across all settings is further italicized.

		1. Recommendation			2. Substance			3. Appropriateness			4. Comparison		
		P ↑	Sp ↑	K ↑	P ↑	Sp ↑	K ↑	P ↑	Sp ↑	K ↑	P ↑	Sp ↑	K ↑
given review	1. zero-shot	**0.826**	**0.836**	**0.757**	0.394	0.414	0.367	0.473	0.489	0.439	0.393	0.399	0.349
	2. few-shot	0.807	0.811	0.733	**0.453**	**0.452**	**0.413**	**0.634**	**0.604**	**0.558**	**0.405**	**0.401**	**0.353**
	3. MCQ style	0.819	0.824	0.744	0.430	0.432	0.382	0.393	0.453	0.392	0.344	0.313	0.273
given paper	4. abstract	**0.283**	**0.282**	**0.25**	**0.187**	**0.190**	**0.166**	0.187	0.152	0.129	-0.023	-0.031	-0.028
	5. whole paper (GPT-3.5-16k)	0.090	0.091	0.080	0.000	-0.003	-0.003	-0.100	-0.019	-0.097	-0.030	0.030	-0.025
	6. selected sections	-0.076	-0.081	-0.072	0.155	0.131	0.117	**0.202**	**0.222**	**0.197**	0.021	0.011	0.009
	7. abstract & sections	-0.014	-0.008	-0.007	0.032	0.051	0.046	0.002	-0.007	-0.007	**0.080**	**0.076**	**0.068**
		5. Soundness			6. Originality			7. Clarity			8. Impact		
		P ↑	Sp ↑	K ↑	P ↑	Sp ↑	K ↑	P ↑	Sp ↑	K ↑	P ↑	Sp ↑	K ↑
given review	1. zero-shot	0.585	0.619	0.542	0.507	0.510	0.443	0.626	0.649	0.572	0.445	0.450	0.389
	2. few-shot	**0.667**	**0.674**	**0.610**	**0.612**	**0.621**	**0.545**	**0.730**	**0.745**	**0.676**	**0.504**	**0.521**	**0.460**
	3. MCQ style	0.398	0.395	0.355	0.476	0.469	0.409	0.726	0.718	0.644	0.494	0.493	0.436
given paper	4. abstract	0.120	0.117	0.104	0.118	0.119	0.098	**0.172**	**0.171**	**0.151**	0.109	**0.113**	**0.097**
	5. whole paper (GPT-3.5-16k)	0.052	0.064	0.057	0.095	0.103	0.085	-0.082	-0.091	-0.081	0.069	0.086	0.071
	6. selected sections	0.008	0.016	0.014	**0.197**	**0.189**	**0.157**	0.081	0.105	0.093	0.088	0.075	0.066
	7. abstract & sections	**0.157**	**0.183**	**0.162**	0.084	0.079	0.067	0.112	0.143	0.128	**0.118**	0.109	**0.097**

Table 2: Detailed aspect prediction results of GPT-3.5 and GPT-3.5-16k on PeerRead dataset. Numbers in gray color are values with p-value larger than 0.05. P is short for Pearson correlation, Sp for Spearman correlation, and K for Kendall's tau.



The exposition is OK, and I think the approach is sensible, but the main issue with this paper is that it is lacking experiments on non-synthetic datasets. As such, while I find the graphics inspired questions the paper is investigating interesting, I don't think it is clear that this work introduces useful machinery for modeling more general videos. I think this paper is more appropriate as a workshop contribution in its current form.

Title: Perception Updating Networks: On Architectural Constraints For Interpretable Video Generative Models **Abstract:** We investigate a neural network architecture and statistical framework that... **Section 1** Introduction: The current computer graphics pipelines... **Section 3** Perception Updating Networks: This Section proposes a family of neural architectures for optimizing...

Recommendation = 2 **Soundness** = 3
Substance = 2 **Originality** = not discussed
Appropriateness = 2 **Clarity** = not discussed
Comparison = not discussed **Impact** = 3

Figure 1: Example of the aspect score prediction task. We conduct experiments under two settings: given the review or the paper to predict scores.

	accuracy ↑	|diff| ↓	Pearson ↑	Spearman ↑	Kendall's tau ↑
most freq	0.317	**0.813**	**0.628**	**0.630**	**0.560**
all 1	0.314	0.952	0.522	0.525	0.459
all 5	**0.337**	0.822	0.546	0.539	0.476

Table 3: Average results of "only given abstract" method on 100 randomly chosen examples from PeerRead dataset. "most freq" means using the most frequent reference score for each aspect in the prompt example; similarly, "all 1" and "all 5" mean setting all scores in the prompt example to 1 and 5.

We conduct experiments under two different settings: (1) given human-written review, predict aspect scores; (2) given (part of) the research paper, predict scores.

3.2. Setting 1: Given Review, Predict Scores

Tests under the Setting 1 can be viewed as a review reading comprehension task, targeting questions: are model-generated scores meaningful? Are they consistent with the text? Can models understand human-written reviews? In addition, we take into consideration the influence of prompt style and content extraction methods. For Setting 1, we try zero-shot / few-shot, direct scoring / multiple-choice style scoring, and different example distributions.

Besides classical metrics *accuracy* and *absolute* *difference* for the score prediction task, we also calculate the correlation indicators of Pearson, Spearman, and Kendall's tau. The three correlation metrics are the common choice when evaluating the score prediction ability of unfinetuned models, as in the work of using LLM to evaluate abstractive summaries (Shen et al., 2023) and machine translations (Kocmi and Federmann, 2023).

If not specially marked, all models are of version 0613 with temperature 0.3, for example GPT-3.5-turbo-0613 in this section.

LLM can infer scores from reviews. As shown in Table 1 (column 3 Pearson), when predicting scores given the review, GPT-3.5 achieves a good correlation with humans (0.651 Pearson correlation value under the few-shot setting, while the baseline is only about 0.3). Even under the most difficult zero-shot setting (line 2 zero-shot & column 345), its correlations are still satisfactory (above 0.5). This indicates that GPT-3.5 can understand human-written reviews, distinguish emotions, and give consistent scores.

Another result worth noticing is that the multiple-choice question style prompting does not help much (line 4 MCQ style). In the MCQ style prompt, we write specific scoring criteria for each score in each aspect but only obtain a small performance gain compared to zero-shot prompting. We may conclude that GPT-3.5 already knows the rules and can inherently give meaningful scores.

3.3. Setting 2: Given Paper, Predict Scores

Experiments under the Setting 2 really assess model's capability to be a reviewer, answering our main research question. Under this task setting, we try three types of input for LLM: only abstract, selected sections, and the whole paper (for GPT-3.5-16k).

LLM fails to predict scores directly from papers. Unlike predicting scores given the review, when only given (part of) the research paper, GPT-3.5 struggles to generate reasonable scores. The bottom half (line 5678) of Table 1 shows that GPT-3.5 only has 0.258 best Pearson correlation, even lower than the baseline. The particularly poor performance of GPT-3.5-16k (line 6) with correlations lower than 0.2 gives us another indication: simply injecting long texts is not the way out, especially complex long texts like research papers.

LLM predicts well the final [recommendation] score, but not scores of [comparison], [impact], and [substance] that are knowledge- and logic-demanding. GPT-3.5 gains especially high correlations in predicting [Recommendation] scores under both settings (Pearson correlation 0.826 & 0.283, see detailed results for each aspect in Table 2 column 1). However, it struggles in judging [Comparison] (column 4) whether the paper presents enough meaningful comparisons with related work, [Substance] (column 2) whether it contains lots of ideas and results, and [Impact] (column 8) whether it is influential and helpful to this field. We may attribute the difficulty to the need of extra scientific knowledge and details, as all three aspects require a rich understanding of the field.

However, we cannot exclude the possibility of using memorized data to successfully predict the [Recommendation] score (data leakage), as this score is the easiest to infer from other factors and the PeerRead dataset uses ICLR-2017 papers. Therefore, we present a more detailed examination of model's ability in Section 5 on our MCQ test data.

We justify the choice of prompt example in Table 3. Using the most frequent score of each aspect in the prompt has the best result, but the influence is not decisive, as the variance among the three prompts' results is small. Therefore, we use the "most frequent" score in the prompt example for all experiments in this section.

4. Task 2: Review Generation

4.1. ASAP dataset

For the task of review generation, we use the ICLR-2020 subset of the ASAP dataset (Yuan et al., 2022). Review sentences in this subset are labeled by their aspect: *summary, motivation, originality, soundness, substance, replicability, meaningful comparison, clarity*; each is further classified into positive and negative, except *summary*. We randomly select 300 papers from this subset with 902 corresponding official peer reviews to test model's review generation ability. An example of GPT-4 generated review is shown in Figure 2.

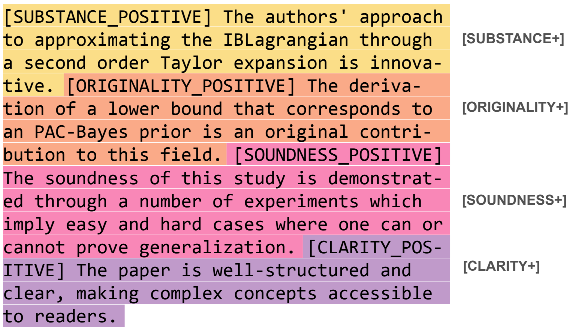

[SUBSTANCE_POSITIVE] The authors' approach to approximating the IBLagrangian through a second order Taylor expansion is innovative. [ORIGINALITY_POSITIVE] The derivation of a lower bound that corresponds to an PAC-Bayes prior is an original contribution to this field. [SOUNDNESS_POSITIVE] The soundness of this study is demonstrated through a number of experiments which imply easy and hard cases where one can or cannot prove generalization. [CLARITY_POSITIVE] The paper is well-structured and clear, making complex concepts accessible to readers.

Figure 2: Example review generated by GPT-4. The sentence aspect label is part of the generation and is put at the beginning of each sentence. A special [None] label is added if the sentence does not belong to any other types.

4.2. Experiments and Results

As in the aspect score prediction test in Section 3, we try zero-shot / few-shot prompting and different content extraction methods in the two-step generation experiment. We present both the performance of GPT-3.5-turbo-0613 and GPT-4-0613 for the review generation task. Besides, we focus more on the evaluation method for model-generated review texts.

We examine the quality from the following perspectives. (1) Aspect coverage, thanks to the aspect annotations in ASAP dataset. (2) Similarity to reference reviews, including the classical statistical methods ROUGE-1/2/L (Lin, 2004), the model-based method BertScore (Zhang et al., 2019), the task-based reference-free metric BLANC (Vasilyev et al., 2020), and using GPT-4 to score the similarity as in the evaluation of abstractive summaries (Shen et al., 2023) and of translation quality (Kocmi and Federmann, 2023). (3) Manual analysis for similarity and informativeness (whether containing enough details) proposed by Yuan and Liu (2022) in KID-Review.

Here are some implementation details. GPT-3.5 is used for 300 randomly selected papers from ASAP, whose results are auto-evaluated in Figure 3 and Table 5; GPT-4 is used for 50 papers and the results are manually examined in Table 6. The reason for only choosing 50 papers for GPT-4 is that, the generation is expensive and that the manual analysis has also a high cost. For the two-step generation experiment, the section extraction step outputs the union of contents that are selected based on each review (most papers have three reference reviews), and the union of useful contents is input into the model to generate one review. During the evaluation, the best similarity score is selected among all references for ROUGE and BertScore; for BLANC, all reference reviews are concatenated to form the new reference.

LLM has its own comment aspect preference: too much positive feedback. In Figure 3, it is clear that GPT-3.5 has its own preference of aspects to comment on: it always generates positive reviews while being very cautious about negative aspects. For each setting, the proportion of positive feedback (exclude [Summary]) is always larger than 55% (reference labels are only 43% positive). Especially when only given the abstract, 99% comments are positive. In addition, [Substance-] and [Clarity-] are always missing in all four experiment settings.

Few-shot prompting helps GPT-3.5 to generate more negative comments, but the distribution is still very different from reality. We regard this as a strong weakness for GPT-3.5 being a reviewer because well-supported critics are the most helpful for researchers.

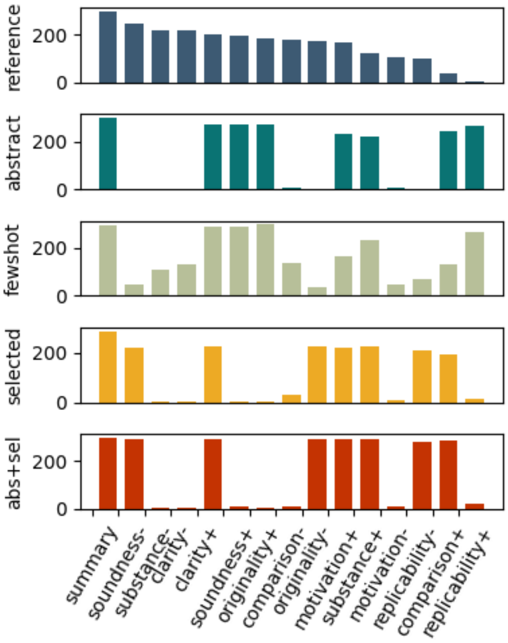


Figure 3: Aspect label distribution of reference reviews and model-generated reviews on ASAP dataset. "+ / -" means positive/ negative comments, except that [summary] is always neutral. Reference label distribution comes from the union of all reviews for each paper.

The aspect recall value in Table 5 (column 1 aspect) also indicates that both GPT-3.5 and GPT-4 have very different comment preferences from humans. Although their best aspect coverage rate (0.582) is better than humans (0.499), the highest aspect recall is still unsatisfactory (0.559). Even under the few-shot prompt setting (line 2 few-shot), this value is still low (0.506), showing that LLM cannot naturally generate comments of people's interest.

Automatic metrics can be fooled by LLM's generation. In Table 4, to better examine the review generation quality, we try using GPT-4 as an evaluator to predict the relevance, precision, and recall score (between 0 and 100). We also manually score (also between 0 and 100) the 50 reviews generated by GPT-4-0613 when given both the abstract and selected sections. We focus on two perspectives for quality evaluation: (1) relevance to reference reviews; (2) amount of details in the review (informativeness).

During the process of manual inspection, we find that GPT-4 generated reviews are strongly influenced by the given paper content and lack critics, while reference reviews contain lots of detailed suggestions. This is why the manual score is low, 54.5 for relevance and 48.7 for informativeness.

However, the manual score (and also our actual im-

GPT-4 predicted	relevance ↑	precision ↑	recall ↑
	83.8	84.7	75.8
manually scored	relevance ↑	informativeness ↑	
	54.5	48.7	

Table 4: "GPT-4 as review quality evaluator" predicted scores and manually annotated scores for the 50 GPT-4 generated reviews.

pression) is very different from BertScore (always above 0.8): is BertScore fooled by the fluent generation? Thus, we compute the correlation between every automatically computed similarity score and our manual annotations in Table 6. As expected, BertScore has the worst correlation (-0.003 and -0.024), showing that BertScore is fooled by GPT-4 generated texts.

Automatic metrics that best correspond to humans are BLANC and GPT-4. BLANC is designed for no-reference summary evaluation, but here we consider the generated review as a summary of all reference reviews. BLANC uses the performance gain on the blank-filling task as the indicator: if a summary can help its model to complete the original task, then the summary is considered of good quality. This metric truly tests the amount of information contained in the summary and is difficult to be fooled by fluent but hollow sentences.

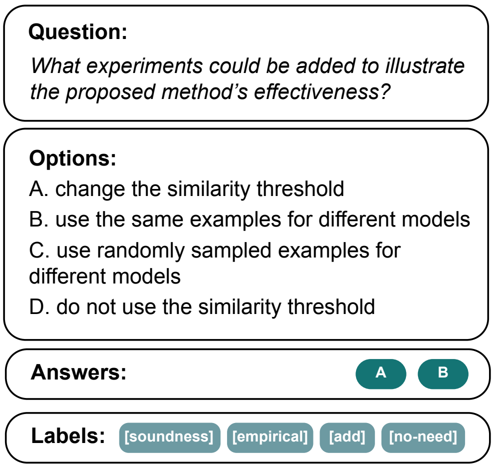
In conclusion, GPT-4 does not naturally provide enough details and critics; automatic similarity metrics can be unaware of the flaws. Since it is still difficult to evaluate model-generated reviews, we use our RR-MCQ data in Section 5, a more objective and detailed approach to evaluation.

5. Task 3: Review-Revision Multiple-Choice Questions

The experiments of aspect score prediction and review generation above do give us an idea of model's reviewing ability. Moreover, GPT-4 has also been tested on the edit generation task in ARIES (D'Arcy et al., 2023). They manually annotate 85 generated examples from the following perspectives: compliance (1-3), promise (true/false), paraphrase (true/false), and technical details (true/false). They find that GPT-4 has high compliance (94% are scored 3), but tends to make promises (21%), simply paraphrase the comment (48%), and lacks technical details (12%).

One evident drawback of the previous analyses is that, all detailed results require heavy manual inspection while automatic metrics are not reliable enough (see Section 4.2). Therefore, we propose a review-revision multiple-choice question dataset (RR-MCQ) with abundant labels, containing questions related to both the review and revision with one or more correct answers.

5.1. RR-MCQ Data Construction and Characteristics



Question:

What experiments could be added to illustrate the proposed method's effectiveness?

Options:

A. change the similarity threshold

B. use the same examples for different models

C. use randomly sampled examples for different models

D. do not use the similarity threshold

Answers: A B

Labels: [soundness] [empirical] [add] [no-need]

Figure 4: Example of the multiple-choice question with one or more answers. We randomly shuffle the four options during the experiment.

Our RR-MCQ dataset is targeted for a more specific and in-depth assessment. For example, can models evaluate the soundness of argumentation? Can they integrate domain knowledge and the paper together? Can they give complicated suggestions, such as important experiments to do?

The dataset contains 196 multiple-choice questions examining specific review-revision-related knowledge and ability. An example of the RR-MCQ is presented in Figure 4.

To construct the MCQ test dataset, we select 55 reviews from 14 papers with sufficient comment-response posts in the peer review forum from the ICLR-2023 conference. We then perform the following four steps: (1) align the smallest unit of comment and response to form a single argument; (2) identify its main topic and decide if controversial (we skip arguments on which the two sides sharply disagree); (3) transform the argument into a four-choice question without adding new contents, i.e. wrong options either come from irrelevant parts of the same discussion or are the negation of correct ones; (4) label the aspects being assessed by the question.

There are four types of labels corresponding to four ways of categorization: review aspect, content aspect, ability to be tested, if need information from other papers. Figure 5 presents the label distribution, which basically corresponds to the review comment distribution in reality. The labels are assigned by two experienced students in the domain. Among all the 788 annotated labels, 86 labels (10.9%) have disagreement at first. The final decision is made through a careful discussion of the two annotators.

		1. aspect (macro avg)			2. ROUGE-F1 (macro avg)		
		ref coverage ↑	coverage ↑	recall ↑	R-1 ↑	R-2 ↑	R-L ↑
abstract	1. zero-shot	0.499	0.034	0.065	0.429	0.103	0.190
	2. few-shot	0.499	**0.582**	0.506	0.424	**0.108**	**0.198**
paper	3. selected sections	0.499	0.367	0.449	0.382	0.081	0.176
	4. abstract & sections	0.499	0.470	**0.559**	**0.431**	0.106	0.193
GPT-4*	5. abstract & sections	0.488*	0.515*	0.530*	0.425*	0.100*	0.190*

		3. BertScore (macro avg)			4. BLANC	
		precision ↑	recall ↑	F1 ↑	average ↑	variance
abstract	1. zero-shot	**0.855**	0.835	**0.843**	0.098	0.001
	2. few-shot	0.851	0.839	**0.843**	0.114	0.001
paper	3. selected sections	0.843	0.832	0.835	0.093	0.001
	4. abstract & sections	0.845	**0.840**	0.841	**0.126**	0.001
GPT-4*	5. abstract & sections	0.851*	0.840*	0.843*	0.112*	0.001*

Table 5: Automatic evaluation of GPT-3.5 (300 examples) and GPT-4 (50 examples) generated reviews on ASAP dataset. GPT-4 results are noted with * because they are averaged over 50 examples.

	relevance	informativeness
aspect-recall	0.076	0.227
Rouge1-F1	0.115	0.159
Rouge2-F1	0.039	0.111
RougeL-F1	0.167	0.124
BertScore-F1	-0.003	-0.024
BLANC-avg	0.255	**0.355**
GPT4-avg	**0.325**	0.244

Table 6: Pearson correlation values between automatic evaluation metrics and manually annotated review quality labels for the 50 GPT-4 generated reviews.

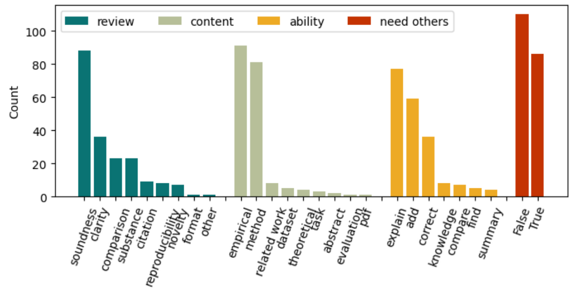


Figure 5: Label distribution of our RR-MCQ test data. There are 4 types of labels: review aspect, content aspect, ability, and if need information from other papers.

5.2. Experiments and Results

We test both GPT-3.5-turbo-0613 and GPT-4-0613 on our MCQ data. The two-step generation method is similar to that of Section 4: the model selects useful sections based on the given question, then the selected contents are input into the model to predict answers. Note that our multiple-choice questions may have more than one correct answer.

LLM gets passable micro accuracy, but bad macro accuracy. From Table 7, the best micro accuracy 0.710 comes from the GPT-4 -> GPT-4 method. It is a passable score, but the macro accuracy is not ideal: the best macro accuracy is only 0.276. The micro accuracy is calculated by considering each question option as an individual example so that the MCQ task becomes a binary decision task of determining True/False for each option. The macro accuracy is more strict: only when all answers are correct, the question is marked as correct (note that the number of correct options is undetermined, between one and four).

In addition, the precision and recall are balanced. We can conclude that GPT-4 is able to judge the correctness of each individual statement with about 70% accuracy, but this level of ability is insufficient for giving completely correct answers: errors and omissions are frequent.

LLM struggles with tasks that relate to soundness and adding components. We select the top 2 numerous labels in each of the 4 label categories and present the detailed results in Table 8. As our questions are inferred from the true peer review discussion forum (containing both the review and the response), these types of questions are also the most common in reality. Therefore, model's performance on these questions is important and representative.

In Table 8, [Explain] related questions have the highest macro accuracy (by comparing the bold number in "accuracy" through all eight aspects). These types of questions come from discussions to confirm an understanding or ask a detail, somehow similar to reading comprehension and essay expansion tasks. It does not require much logi-

	macro average				micro average			
	accuracy ↑	precision ↑	recall ↑	F1 ↑	accuracy ↑	precision ↑	recall ↑	F1 ↑
GPT-3.5 -> GPT-3.5	0.128	0.332	0.376	0.176	0.569	0.583	0.373	0.227
GPT-4 -> GPT-3.5	0.214	0.553	0.586	0.285	0.648	0.644	0.603	0.311
GPT-4 -> GPT-4	**0.276**	**0.655**	**0.666**	**0.330**	**0.710**	**0.699**	**0.701**	**0.350**

Table 7: Results on RR-MCQ test data. “GPT-4 -> GPT-3.5” means using GPT-4 for the first section selection step and then using GPT-3.5 for the second answer generation step.

	1. Soundness			2. Clarity			3. Empirical			4. Method		
macro average	acc ↑	prec ↑	recall ↑	acc ↑	prec ↑	recall ↑	acc ↑	prec ↑	recall ↑	acc ↑	prec ↑	recall ↑
GPT-3.5 -> GPT-3.5	0.091	0.293	0.307	0.194	0.356	0.454	0.055	0.267	0.308	0.173	0.391	0.458
GPT-4 -> GPT-3.5	**0.205**	0.526	0.552	0.278	0.498	**0.560**	0.209	0.573	0.597	0.247	0.564	0.598
GPT-4 -> GPT-4	0.193	**0.655**	**0.673**	**0.361**	**0.514**	0.509	**0.253**	**0.713**	**0.695**	**0.309**	**0.617**	**0.654**
	5. Explain			6. Add			7. No need			8. Need		
macro average	acc ↑	prec ↑	recall ↑	acc ↑	prec ↑	recall ↑	acc ↑	prec ↑	recall ↑	acc ↑	prec ↑	recall ↑
GPT-3.5 -> GPT-3.5	0.182	0.374	0.384	0.051	0.284	0.336	0.118	0.295	0.358	0.140	0.380	0.400
GPT-4 -> GPT-3.5	0.247	0.573	0.534	**0.203**	0.586	0.609	0.227	0.523	0.549	0.198	0.591	0.634
GPT-4 -> GPT-4	**0.364**	**0.667**	**0.626**	0.153	**0.694**	**0.757**	**0.291**	**0.602**	**0.647**	**0.256**	**0.724**	**0.691**

Table 8: Detailed results on RR-MCQ test data. For each label category, we present only the detailed results for the top two labels with the most elements.

cal reasoning, but lots of knowledge and inference ability.

Question [Soundness]:

Which choice the author makes may hurt the final performance but is not well-examined in the paper?

Options:

A. linearize the RNN module

B. use a learnable gate

C. combine RNN and self-attention

D. REM endows positional encodings of a multi-head self-attention with recurrent dynamics

Answers: A

Figure 6: Example question of label [Soundness].

The lowest macro accuracy appears in questions of labels [Soundness] and [Add] (note that one question may have both the [Soundness] and [Add] labels because they mark different perspectives).

An example of [Soundness] question is shown in Figure 6. The review aspect label [Soundness] concerns questions requiring strong logic, for example, the correctness of a statement, the validity of an argument, or the completeness of supporting evidence. In the example above, the model needs to first identify whether the choice is influential, and then decide if it is well-examined in the paper.

An example of [Add] question is shown in Figure 7.

Question [Add]:

What possible experiments can be added to present the method’s usefulness ?

Options:

A. compare to linear freezing and AutoFreeze

B. try DeiT-T model on ImageNet dataset

C. try Bert model on MRPC dataset

D. try Bert model on CoLA dataset

Answers: B C D

Figure 7: Example question of label [Add].

The tested ability label [Add] relates to adding components to the paper, for example, conducting an extra experiment or citing a missing related work. Its difficulty comes from the need for both logic and knowledge. Although possible options to be added are already provided in the question, the model still has to carefully select truly necessary ones. In the example question above, GPT-4 fails to understand that the first option is already in the paper and that the second option is needed to prove method’s effectiveness on other model architectures.

6. Conclusion

“Can LLM be a qualified and reliable automatic reviewer?” After testing GPT-3.5 and GPT-4 on two existing datasets and also on our proposed RR-MCQ data, we conclude that they are not naturally reliable automatic reviewers because their error rate is still not sufficiently low.

They can generate meaningful scores based on human-written reviews, even without explicitly giving examples or scoring criteria; but when the input is long and complicated like a whole research paper, they can only roughly identify the quality. When being asked to freely generate comments, their suggestions are sometimes correct, but always on aspects that human reviewers would not be interested in. Facing multiple-choice questions, they have the ability to make passable decisions on single options, but hardly to be completely correct .

We claim that it is still too early to trust LLM as automatic scientific paper reviewer. Although there is a chance to get useful and correct results, their current capability is not reliable enough. Especially on questions requiring logic reasoning over long texts or extra knowledge in detail, their performance is still unsatisfactory.

We believe that in detail and in-depth evaluations are needed before the targeted development of LLM in automatic paper reviewing task. Our RR-MCQ test dataset is an example, but still with limited size. Future work could be to develop better data construction methods, to invent new metrics, and finally to improve LLM's capability as a reviewer.

7. Acknowledgements

This work was supported by the National Key R&D Program of China 2023ZD0120703 and the China NSFC Projects (U23B2057, 62106142 and 62120106006) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

8. Bibliographical References

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

Gautam Choudhary, Natwar Modani, and Nitish Maurya. 2021. React: A re view comment dataset for act ionability (and more). In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part II 22*, pages 336–343. Springer.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *arXiv preprint arXiv:2211.06651*.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4758–4766.

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one*, 17(1):e0259238.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. *arXiv preprint arXiv:1903.10104*.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arxivedits: Understanding the human revision process in scientific writing. *arXiv preprint arXiv:2210.15067*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2021. Disapere: A dataset for discourse structure in peer review discussions. *arXiv preprint arXiv:2110.08520*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. Multi-task peer-review score prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 121–126.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, pages 1–16.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

OpenAI. 2023. Gpt-4 technical report.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2021. Mred: A meta-review dataset for structure-controllable text generation. *arXiv preprint arXiv:2110.07474*.

Shruti Singh, Mayank Singh, and Pawan Goyal. 2021. Compare: a taxonomy and dataset of comparison discussions in peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 238–241. IEEE.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*.

Weizhe Yuan and Pengfei Liu. 2022. Kid-review: Knowledge-guided scientific review generation with oracle pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11639–11647.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

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

A. Prompt

A.1. Evaluation on PeerRead

Setting 1 Given review, predict scores.

Prompt You are a professional reviewer in computer science and machine learning. Based on the given review, you need to predict the review score in several aspects. Choose a score from [1,2,3,4,5], higher score means better paper quality.

Zero-Shot Example Example output: RECOMMENDATION: x, SUBSTANCE: x, APPROPRIATENESS: x, MEANINGFU COMPARISON: x, SOUNDNESS CORRECTNESS: x, ORIGINALITY: x, CLARITY: x, IMPACT: x

Few-Shot Example Example1: Review: This paper presents an approach to modeling videos based on a decomposition into a background ... workshop contribution in its current form. Output: RECOMMENDATION: 2, SUBSTANCE: 2, APPROPRIATENESS: 2, SOUNDNESS CORRECTNESS: 3, IMPACT: 3 . Example2 ... Example5 ...

MCQ-Style Example RECOMMENDATION: A. This paper changed my thinking on this topic and I'd fight to get it accepted; B. I learned a lot from this paper and would like to see it accepted. C. Borderline: I am ambivalent about this one. D. Leaning against: I would rather not see it in the conference. E. Poor: I would fight to have it rejected.

Setting 2 Given paper, predict scores.

Prompt You are a professional reviewer in computer science and machine learning. Based on the given abstract/sections/paper, you need to predict the review score in several aspects. Choose a score from [1,2,3,4,5], higher score means better paper quality.

A.2. Evaluation on ASAP

Setting 1 Given paper, generate review text.

Prompt You are a professional reviewer in computer science and machine learning. Based on the given title and abstract of a research paper, you need to write a review in ICLR style. At the same time, you need to tag sequences of words with their review type at the beginning: [NONE], [SUMMARY], [MOTIVATION POSITIVE], [[MOTIVATION NEGATIVE]], [SUBSTANCE POSITIVE], [SUBSTANCE NEGATIVE], [ORIGINALITY POSITIVE], [ORIGINALITY NEGATIVE], [SOUNDNESS POSITIVE], [SOUNDNESS NEGATIVE], [CLARITY POSITIVE], [CLARITY NEGATIVE], [REPLICABILITY POSITIVE], [REPLICABILIT NEGATIVE], [MEANINGFUL COMPARISON POSITIVE], [MEANINGFUL COMPARISON NEGATIVE]. Your total output should not surpass 500 tokens.

Zero-Shot Example Example output: [LABEL] sequence. [LABEL] sequence. [LABEL] sequence......

Few-Shot Example Example1: [SUMMARY]This paper introduces a method to disentanglement the private and public attribute information... Example2: [SUMMARY]The paper proposes learning NN to correct for inaccuracies... Example3: [SUMMARY]This paper describes a method for segmenting 3D point clouds... Example4: [SUMMARY]This work introduces GQ-Net , a novel technique that trains quantization friendly networks...

Setting 2 Given reference reviews, evaluate the generated review quality.

Prompt Score the following review step by step with respect to its relevance with reference reviews on a continuous scale from 0 to 100. You should give a relevance score, a precision score and a recall score of the review to be scored. Relevance measures its selection of important content from references, where relevance=0 means 'no meaning preserved' and relevance=100 means 'perfect meaning'. Precision measures its correctness with respect to references, and recall measures its information coverage with respect to references. Output format: Score for the review to be scored:relevance=x, precision=x, recall=x.

A.3. Evaluation on RR-MCQ

Setting 1 Given question and paper, select useful sections.

Prompt You are a professional reviewer in leywords. You will be given a multiple choice question and the headings of a research paper in this field. You need to select sections that are useful to anwer the question.

Setting 2 Given selected sections, predict answers.

Prompt You are a professional reviewer in keywords. You will be given some sections extracted from a paper in this domain. Based on the given context, you need to answer the following multiple choice question. You should select one or more answer choices from A, B, C, D.

B. Labeling Principle

B.1. Review aspect

- [Soundness] Questions related to the soundness of claims, supporting materials and mathematical results.
- [Clarity] Questions of requesting more explanations.
- [Comparison] Questions related to the comparison with related work: whether it is precise and complete.
- [Substance] Questions to evaluate the number of new ideas, results and the amount of work.
- [Citation] Specific questions of citations without much comparison.
- [Reproducibility] Questions about code availability, settings and hyperparameters.
- [Novelty] Questions to evaluate the significance of problem, technique, methodology, or insight.
- [Format] Specific questions about the paper format.

B.2. Content aspect

Questions related to different parts of the paper. The labels are: [Empirical Result], [Method], [Related Work], [Dataset], [Theoretical Result], [Task], [Abstract], [Evaluation], [PDF].

B.3. Ability

The main ability needed to solve the question. If more than one ability is required, only choose the more complex one. The following labels are ordered increasingly by their complexity.

- [Knowledge] Questions about domain knowledge, not the paper.
- [Summarize] Questions about general descriptions of the paper. If it requires detailed information or analysis of the reasoning process, the use the [find] label.
- [Compare] Questions about comparing the paper to other domain knowledge.
- [Find] Questions about detailed information and logic.
- [Explain] Questions about further explanations. If the explanation needs to add content, for example extra experiments or results, then use the [add] label.

- [Add] Questions about adding content to the original paper. For example experiments, results, citations, etc. If the correction of old content is also involved, then use the [correct] label.
- [Correct] Questions about finding errors and make modifications.

B.4. Extra Information

Only information from referenced papers (citations in the paper or in the discussion forum) are considered extra information.