EXPLORING CROWD SOURCED WORKER EVALUATION METHODS IN OPEN-ENDED TASKS

Ryuya Itano[1], Honoka Tanitsu[1], Motoki Bamba[1], Ryota Noseyama[2], Akihito Kohiga[2], and TakahiroKoita[1]

[1] Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan
[2] Faculty of Science and Engineering, Doshisha University, Kyoto, Japan

ABSTRACT

Crowd sourcing assumes a transient relationship between task requesters and workers, which makes it hard for workers to improve their skills. In addition, crowd sourcing tasks are shifting from simple to more complex and open-ended, highlighting the importance of training workers to handle such tasks. Although various methods have been proposed to train workers, a method to evaluate their skill levels in open-ended tasks has not yet been established. Direct evaluation by requesters is desirable, but scaling up tasks is difficult due to the requester's heavy workload. This study aims to explore methods for evaluating workers without increasing requesters' workload, comparing and verifying a peer-based method and an LLM-based automated method. The experiment investigated the alignment between evaluations from the two methods and those from requesters, thereby clarifying the characteristics of each method. The experimental results demonstrated the applicability of LLMs to evaluating workers in open-ended tasks, revealing both their strengths in consistency and limitations in capturing subtle human judgments.

KEYWORDS

Crowd sourcing, Worker Training, Worker Evaluation, Amazon Mechanical Turk, LLM-as-a-Judge

1. INTRODUCTION

Crowd sourcing is a work model that outsources various tasks to a decentralized and unspecified large group of people (workers) in exchange for payment. Crowd sourcing has been widely adopted due to its advantages for both task requesters and workers: For task requesters, crowd sourcing allows them to outsource a large volume of tasks at a low cost. For workers, it enables them to work flexibly without time or location restrictions.

Although these advantages have promoted the development of the crowd sourcing working model, the model faces inherent challenges. In this model, workers are treated as a temporary workforce. This leads to inadequate evaluation of workers' contributions, which reduces their work engagement and slows skill improvement.Low work engagement discourages workers from producing consistent high-quality outputs or investing in skill improvement, threatening the long-term reliability and productivity of crowd sourcing.

Traditionally, crowd sourcing has assumed workers are frequently replaceable. To maintain quality under these conditions, existing methods have relied on worker redundancy such as aggregating outputs through majority voting [1].In addition, injecting ground truth (known correct answers) to measure worker performance has been widely adopted [2].These methods work well when a task has a unique correct answer and ground truth is available. However, the

importance of crowd sourcing tasks is shifting from simple classification to more complex and open-ended generation tasks, making it difficult to rely on worker redundancy or ground truth [3].Open-ended tasks are difficult to solve even with many workers if the worker group is insufficiently experienced. Under such conditions, requesters should focus on training and retaining skilled workers capable of solving open-ended tasks.

In open-ended tasks, where numerous valid answers are possible, workers are expected to produce outputs that align with the requesters' expectations. Therefore, training workers to perform well in open-ended tasks demands evaluation and feedback from requesters. Providing direct feedback from requesters has been shown to improve workers' output quality[4].Furthermore, corrective feedback from requesters has been shown to facilitate workers' understanding of requesters' expectations [5].However, providing direct evaluation or feedback to workers places a heavy workload on requesters, making crowd sourced tasks difficult to scale up.

As detailed in the next section, for evaluating worker outputs without increasing requesters' workload, two main methods can be cited: the peer-based method where workers evaluate each other, and the automated method where machines perform evaluations automatically. The peer-based method has concerns about inconsistency and generosity in judgments, while the automated method faces challenges in capturing requesters' subtle intentions. The validity of evaluations impacts workers' acceptance of tasks, motivation for skill improvement, and work engagement. Therefore, investigating the alignment of the evaluations from each method with those from requesters is important.

In this study, we focus on worker evaluation as a preliminary stage of worker training. The objective of this study is to explore worker evaluation methods that align with evaluations from requesters in open-ended tasks without increasing requesters' workload. The main contributions of this study are as follows:

- We empirically compared two worker evaluation methods—the peer-based method and the automated method—in open-ended tasks.
- We clarified the characteristics of each evaluation method and demonstrated the importance of selecting methods according to evaluation criteria.
- We demonstrated the applicability of LLMs to evaluating workers in open-ended tasks, revealing both their strengths in consistency and limitations in capturing subtle human judgments.

2. RELATED WORK

This section introduces two main methods for evaluating workers without increasing requesters' workload: the peer-based method, where workers evaluate each other, and the automated method using rule-based models, deep learning models, and large language models (LLMs).

2.1. Peer-Based Method

Peer-based methods such as peer feedback offer a scalable way to support worker skill improvement without direct requester involvement. Several studies have shown that peer feedback among workers contributes to workers' skill improvement and higher-quality outputs [6, 7]. Also, not only receiving but also providing feedback enhances workers' skills [8, 9]. However, few studies have examined the alignment between peer-based evaluations and evaluations from requesters. In the field of education, a study has demonstrated that peer assessment among

students produces evaluations sufficiently consistent with instructor evaluations, reducing instructors' workload despite some variation in evaluations from students[10].

In such decentralized peer evaluation systems, workers tend to give generous evaluations. A study has proposed an approach that organizes workers into a guild-like hierarchical structure where peer evaluations are conducted across different levels[11].This hierarchical organization was shown to suppress generous evaluations from workers. Moreover, a study of data work platforms in China indicates that guild-like hierarchical organizations help to enhance workers' commitment and work engagement [12].The hierarchical structure suggests that workers feel more accountable because their evaluations determine their peers' standing, thereby reducing generous evaluations. However, previous studies have only examined the alignment between workers' evaluations and the ground truth in tasks with unique correct answers. Therefore, we applied the hierarchical peer evaluation structure to open-ended tasks where no unique correct answer exists [13].The hierarchical structure we applied is shown in Figure 1. This structure is constructed by classifying workers into levels according to their skills and is used for the peer evaluation of their outputs. Each worker evaluates those in the level below. The characteristics of this hierarchical peer evaluation structure warrant further verification.

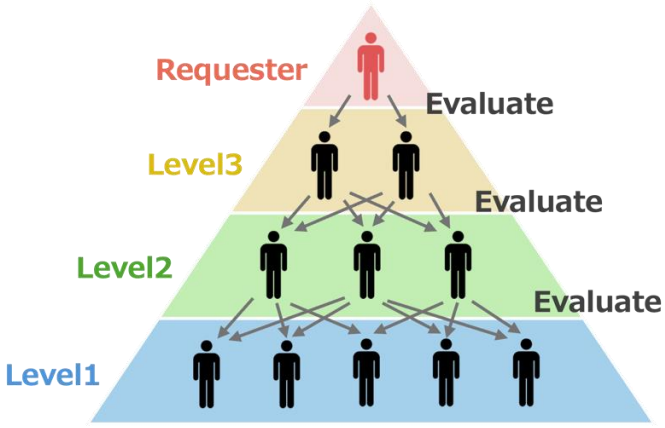

Figure 1. Hierarchical Peer Evaluation Structure.

2.2. Automated Method

Automated methods without direct human involvement have also been considered. In the open-ended tasks of image captioning, one study attempted to automatically evaluate image captions generated by workers [14].The study evaluated captions using rule-based or deep-learning-based natural language processing. The results showed that while the amount of detail in generated sentences aligned closely with human evaluations, criteria such as correctness and fluency remained difficult to evaluate automatically in alignment with human judgment.

Using LLMs for human imitation has also been attempted. In a prior study, LLMs demonstrated the potential to imitate general human political and social attitudes, but they have also exhibited limitations in capturing subtle nuances and context-dependent aspects [15].Whether LLMs can effectively serve as evaluators that reflect specific human subjective criteria remains to be investigated.

The concept of "LLM-as-a-Judge," which utilizes LLMs as evaluators, hasrecently been proposed.A study evaluated AI-generated counter-narratives to hate speech using LLMs and

examined how closely these evaluations aligned with those from humans [16].The results showed that LLMs can evaluate AI-generated text nearly as well as humans(achieving a 0.82 correlation)when quality criteria are explicitly defined, outperforming traditional metrics such as BLEU.These findings suggest the potential for applying LLMs to the evaluation of workers performingopen-ended tasks, which warrants empirical investigation.

3. EXPERIMENT

This section describes the experimental procedure. Briefly, the process involves first aggregating outputs from workers in open-ended tasks, followed by a manual evaluation of these outputs. Subsequently, evaluations are collected using both peer-based and automated methods, and the results are analyzed.

3.1. Task Selection

In this experiment, we adopted an image captioning task as an open-ended task, following a previous study[14].That study included a human-evaluation process for captions generated by workers. The captions were intended to be used for visually impaired individuals and were evaluated on the following three criteria:

- **Fluency**: The smoothness and readability of the text.
- **Correctness**: The factual accuracy of the caption and its relevance to the image.
- **Amount of Detail (Detail)**: The thoroughness with which key visual elements are described.

We adopted this task because the subjective human evaluation results are quantified using these three criteria, facilitating a numerical comparison of the results. For visually impaired individuals, fluency, correctness, and detail are all simultaneously essential. These criteria are mutually independent, and a deficiency in any one of them can reduce the usefulness of a caption.

3.2. Caption Aggregation

We recruited 30 workers from Amazon Mechanical Turk (*https://www.mturk.com*), a crowd sourcing platform, to generate captions for six images. We used six images per worker to evaluate workers' general abilities rather than their performance on individual images, as a smaller number of images would lead to high variability. The six images were selected from the MS COCO captions dataset (*https://cocodataset.org*) following the prior study [14].Figure 2 shows part of the captioning task. The reward for the task was set at $0.50 per worker for completing the entire set of images. Although the number of participants was limited due to evaluation workload constraints, this experiment was designed to serve as an initial validation.

3.3. Evaluation from Requesters

We collected evaluations from requesters to serve as the ground truth for comparison. Acting as requesters, we manually assigned three scores (fluency, correctness, and detail) to each of the six captions produced by every worker on a scale of 0 to 10. Of the 30 workers, the captions from 22 were valid for evaluation. The remaining eight workers were excluded due to duplicated answers or incomplete submissions.

3.4. Evaluation via Peer-Based Method

We used a hierarchical peer evaluation structure as the implementation of the peer-based method. We assigned skill levels to the 22 workers based on the requesters' evaluations. The workers were divided into three levels based on the overall average of their scores across all six image captions, ranked in descending order. The scores were averaged across the three criteria to provide a composite measure of each worker's ability to produce captions that align with the requesters' intent. As shown in Figure 3, the top four workers were assigned to Level 3 (L3), the next six to Level 2 (L2), and the remaining twelve to Level 1 (L1).

Then, we prepared a task to evaluate image captions produced by L2 and L1 workers. Since the L3 workers were supposed to be evaluated directly by requesters, their captions were not evaluated by other workers. Consequently, eighteen workers from L1 or L2 were subject to peer evaluation. The evaluation task was designed so that each worker in L1 and L2 was evaluated by two workers from the immediately superior level (i.e., L3 evaluating L2, and L2 evaluating L1). The evaluation task required assigning three scores (fluency, correctness, and detail) to each of the six captions on a scale of 0 to 10, consistent with the requesters' evaluation process. The reward for the evaluation task was set at $0.80 per worker. Figure 4 shows part of the evaluation task. Ultimately, evaluation scores for fourteen workers were collected. Since several workers did not perform the evaluation tasks, scores for the full set of 18 workers could not be collected as planned.

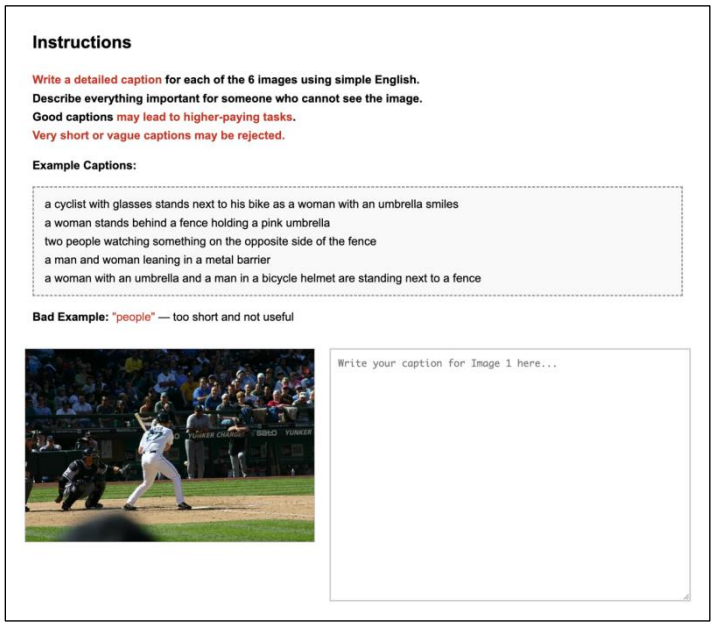

Figure 2. Interface for the Captioning Task

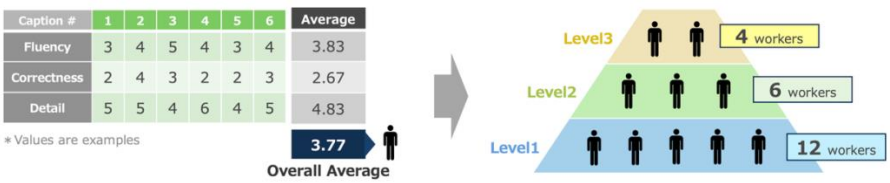

Figure 3. Worker Level Assignment Process

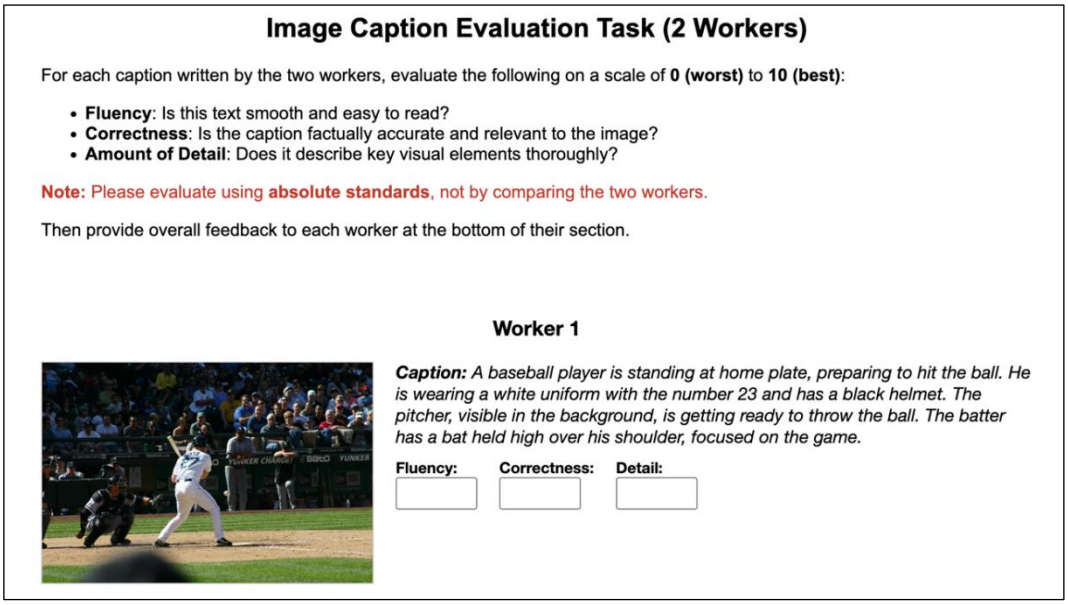

Figure 4. Interface for the Evaluation Task

3.5. Evaluation via Automated Method

We used GPT-5.2 by OpenAI (*https://openai.com*) as the implementation of the automated method. While fourteen workers received evaluations via the peer-based method in the previous step, this step involved obtaining evaluations from GPT-5.2for the six image captions per worker, totaling 84captions.For each evaluation, the prompt shown in Figure 5 and the corresponding image were provided as multimodal inputs to GPT-5.2.To ensure the robustness and reliability of the evaluations, following a prior study [17], we configured GPT-5.2 to output the evaluation results in JSON format, including scores from 0 to 10 for the three criteria along with their reason. To reduce variability, the outputs were generated five times and then averaged.

```
You are given an image and its caption.
Evaluate the caption on a scale of 0 (worst) to 10 (best) based on each of
the three criteria below and provide a reason for the scores.

[Criteria]
1. Fluency: Is this text smooth and easy to read?
2. Correctness: Is the caption factually accurate and relevant to the image?
3. Amount of Detail: Does it describe key visual elements thoroughly?

[Image Caption]
{caption}
```

Figure 5. Input Prompt forGPT-5.2 (The actual caption fills the {caption} field.)

3.6. Analysis

Through the above processes, fourteen workers were assigned scores on three criteria (fluency, correctness, and detail) using three different methods: evaluations from requesters (requester scores), evaluations via the peer-based method (peer scores), and evaluations via the automated method (LLM scores). Scores for each caption were averaged to evaluate workers' overall

capabilities rather than their performance on individual images. Consequently, each worker was assigned nine scores in total: three criteria across three evaluation methods.

First, we calculated the standard deviations (SDs) of the scores across the fourteen workers to analyze the variability within each evaluation method. Specifically, we computed the SDs for each of the nine combinations: three criteria (fluency, correctness, and detail) across three evaluation methods (requester, peer, and LLM scores).We also calculated the following metrics between requester scores and peer scores, as well as between requester scores and LLM scores, to characterize their alignment from complementary perspectives:

- **Mean Absolute Error (MAE)**: Typical magnitude of differences from requester scores.
- **Mean Bias**: Systematic tendency for scores to be higher or lower than requester scores.
- **Pearson Correlation Coefficient (r)**: Linear association with requester scores.
- **Spearman Correlation Coefficient (ρ)**: Agreement in worker ranking with requester scores.

4. RESULTS

This section presents the results of the experiment described previously. First, we report the scores and their corresponding SDs for each evaluation method. Subsequently, we present the comparative metrics assessing the alignment between requester scores and peer scores, as well as between requester scores and LLM scores.

4.1. Scores from Each Evaluation Method

Figure 6 presents line graphs of fluency scores across the evaluation methods, plotted against the worker index. Figures 7 and 8 present the correctness and detail scores, respectively (see Table A in the Appendix for individual worker scores). The worker index is arranged in ascending order based on the requester scores. The dashed lines at the top and bottom of the graph represent the maximum and minimum values for each evaluation method. Table 1reportsthe SDs of the scores from each evaluation methodacross the three criteria.

4.2. Comparative Metrics

Tables 2 and 3 report the comparative metrics between requester and peer scores, as well as between requester and LLM scores, respectively. The mean bias indicates the direction and magnitude of the deviation from the requester scores, where positive or negative values signify systematic overestimation or underestimation.

4.3. Relationship between Caption Length and Detail Score

We additionally examined the relationship between the average length of the captions generated by each worker and the corresponding detail scores across the three evaluation methods. Table 4 reports the Pearson correlation coefficients (r) between caption lengths and detail scores for each method. Figure 9 shows scatter plots of the detail scores for each method plotted against caption length.

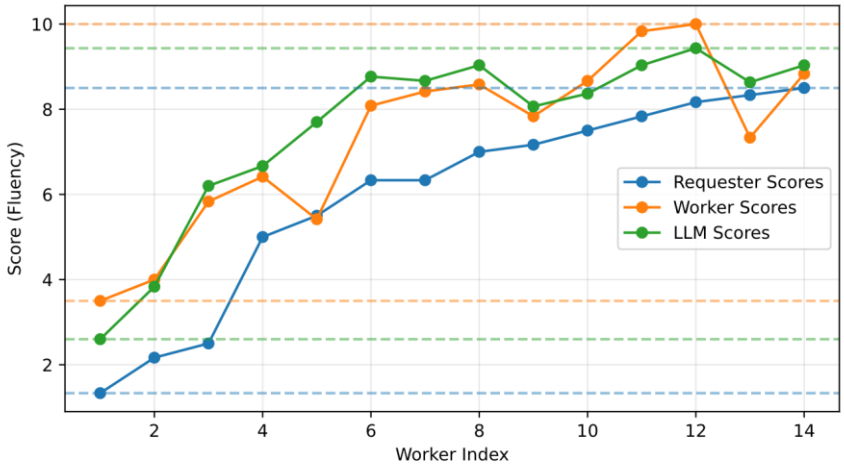

Figure 6. Fluency Scores from Each Evaluation Method

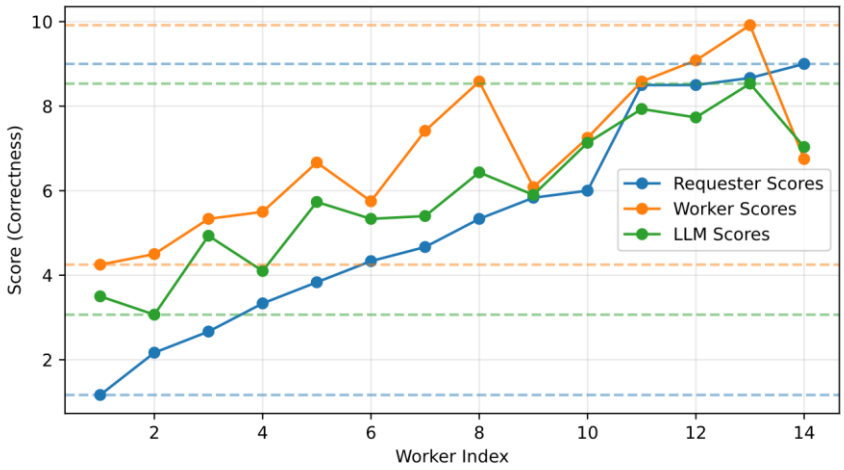

Figure 7. Correctness Scores from Each Evaluation Method

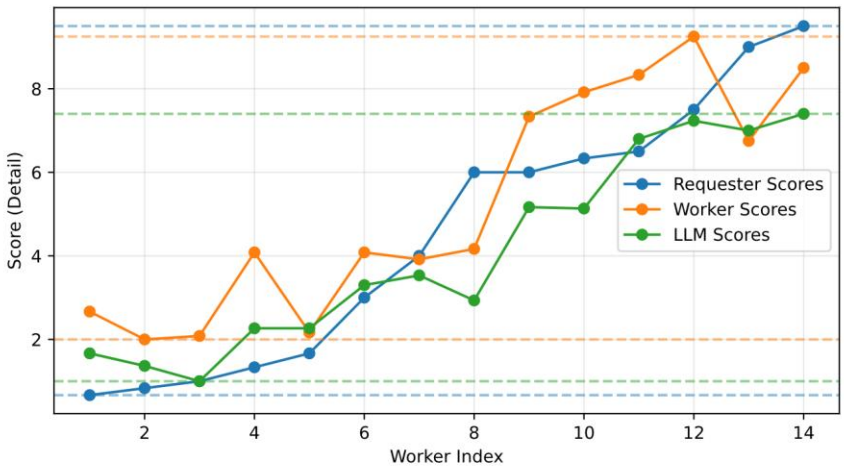

Figure 8. Detail Scores from Each Evaluation Method

Table 1. Standard Deviation of Scores for Each Evaluation Methodacross Three Criteria

Criterion	Requester Scores	Peer Scores	LLM Scores
Fluency	2.395	2.029	2.074
Correctness	2.594	1.735	1.666
Detail	3.126	2.663	2.331

Table 2. Comparative Metrics between Requester and **Peer** Scores

Criterion	MAE	Mean Bias	Pearson r	Spearmanρ
Fluency	1.518	+1.363	0.892	0.814
Correctness	1.869	+1.548	0.826	0.834
Detail	1.446	+0.708	0.881	0.888

Table 3. Comparative Metrics between Requester and **LLM** Scores

Criterion	MAE	Mean Bias	Pearson r	Spearmanρ
Fluency	1.598	+1.598	0.930	0.809
Correctness	1.117	+0.626	0.931	0.928
Detail	0.971	-0.448	0.937	0.952

Table 4. Correlations between Caption Lengths and Detail Scores for Each Method

Requester Scores	Peer Scores	LLM Scores
0.932	0.789	0.795

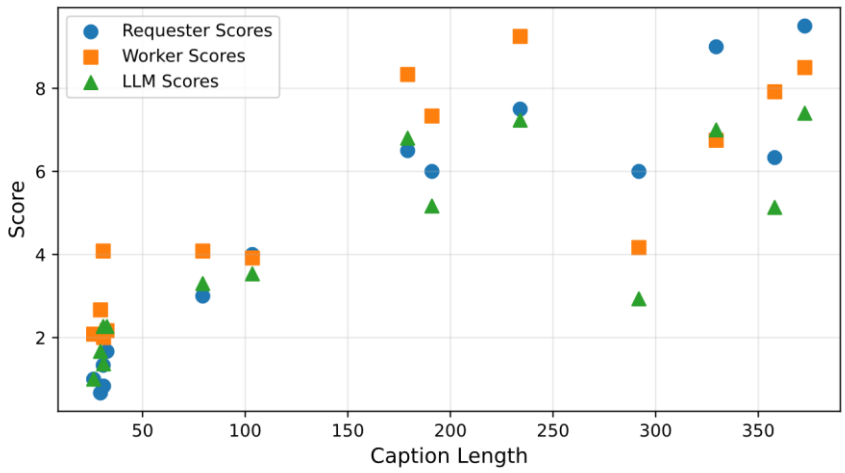

Figure 9. Detail Scores for Each Method against Caption Length

5. DISCUSSION

5.1. Evaluation Variability

As shown in Table 1, the requester scores exhibit the highest SDs across all criteria, suggesting that requesters hold diverse perspectives and expectations. In contrast, the lower SDs observed in

the peer and LLM scores indicate that individual preferences are less likely to be reflected in these evaluations. Notably, the detail criterion showed the highest SDs across all evaluation methods; this suggests that, interestingly, the criterion most likely to align with objective metrics, such as sentence length, is the least consistent in subjective evaluation.

Comparing peer scores and LLM scores, the SDs for LLM scores were slightly higher for fluency, but for the other criteria, the SDs were higher for peer scores. While humans can naturally assess fluency, this may be more challenging for LLMs, potentially leading to greater variation in their interpretation of the fluency criterion. Nevertheless, as the difference is marginal, the two methods can be considered nearly equivalent in this regard. Conversely, the SDs of peer scores exceeded those of LLM scores for the criteria of correctness and detail. This suggests that while human evaluators may sacrifice consistency due to personal preferences, they may be better equipped to capture more subtle nuances.

5.2. Alignment with Requester Scores

As shown in Tables 2 and 3, the peer scores exhibited a lower MAE than the LLM scores in fluency, indicating that humans are better equipped to reproduce the ability to detect subtle naturalness in sentences held by requesters. Furthermore, the MAEs for correctness and detail were lower for the LLM scores, which is likely due to the variability in human evaluations compared to the superior accuracy and consistency of LLMs.

Regarding the mean bias, the peer scores showed positive values across all criteria, indicating that workers tended to evaluate their peers more generously. This finding contradicts those from a prior study suggesting that hierarchical peer evaluation structures can overcome generous scoring [10].However, in our experiment, skill levels were not linked to incentives such as pay raises, which may have failed to foster a sufficient sense of responsibility among workers regarding the evaluation process.

In terms of fluency, the LLM scores showed a higher mean bias than the peer scores, suggesting that the LLM's criteria for fluency are more generous than those of humans. The captions may have been overrated by LLMs because of its emphasis on grammatical and structural accuracy, which led them to overlook subtle unnaturalness perceived by human evaluators. Regarding correctness, the LLM scores showed a significantly lower mean bias than those of the peer scores. As discussed in the MAE analysis, this is attributed to the LLMs' superior ability to detect factual inaccuracies. For the detail criterion, the LLM scores were the only ones to exhibit a negative mean bias, the reasons for which are explained in the next subsection.

Regarding thePearson $r$and Spearman ρ correlations, while the peer scores were consistently strong at around 0.8 across all criteria, the LLM scores performed even better with values frequently reaching the 0.9 range. In the fluency evaluations by LLMs, a notable discrepancy was observed between the Spearman ρ (0.809), and Pearson r(0.930).With a small sample size of $n = 14$, the alignment at both the low and high score boundaries may have pushed up the Pearson r value, potentially overestimating the alignment with the requester scores.

Based on these results, automated evaluation method using LLMs can be considered generally superior to the peer-based method. Particularly in terms of correctness and detail, LLM scores exhibited a clear alignment with the requester scores. However, regarding the assessment of linguistic naturalness, such as fluency, human evaluators demonstrated a superior ability to discern finer nuances. This suggests the importance of selecting the appropriate evaluation method based on the specific criteria. In the future, a hybrid approach, where LLMs serve as the

primary evaluators while human evaluators are utilized for criteria requiring nuanced judgement, is likely to be effective.

5.3. Analysis of Detail Scores

According to Figure 9 and Table 4, the requester scores in this experiment exhibited the strongest correlation with caption length. This indicates that the requester's tendency to favor longer captions was clearly reflected in their evaluations.

The scatter plots in Figure 9 show that the LLM scores do not increase linearly with caption length but instead exhibit a nonlinear pattern. This characteristic resulted in a negative mean bias. A previous study had pointed out a verbosity bias in models prior to GPT-4, where longer texts were preferred [18]. However, the GPT-5.2 model used in this study demonstrated an improvement in mitigating this bias, revealing that it even penalizes longer text. Furthermore, the scatter plots suggest that the high variation in peer scores is the primary cause of their lower correlation with caption length.

Since a requester's intent can vary based on subjective preferences, such as the belief that "excessively long text is undesirable," it is necessary to modify the prompt to explicitly state these preferences when using LLMs. However, for more complex and abstract tasks, the evaluation criteria become increasingly ambiguous and difficult to state. Therefore, further verification is required to explore whether evaluations using LLMs can align with requester intentions in more complex scenarios.

5.4. Limitations

A primary limitation of this study is the small sample size. Evaluations were obtained for only fourteen workers, which is smaller than initially planned. Due to the limited sample size($n = 14$), the results should be interpreted as exploratory.The unexpectedly high attrition rate of workers who did not participate in the evaluation task highlights the challenge of low worker retention in crowd sourcing. Regardless of the type of task, maintaining high worker retention remains an ongoing challenge.

In addition, this study focused solely on the image captioning task, which represents one type of open-ended task. Consequently, the observed superiority of LLMs in evaluating image captions may not necessarily extend to the evaluation of workers performing more complex open-ended tasks.

6. FUTURE WORK

We plan to expand the scale of our experiments to a size sufficient for robust statistical analysis and to verify our findings across other more complex tasks including the evaluation of creative content, ethical judgments, and context-dependent decisions.

While this study focused on methods for evaluating worker skills, we plan to utilize these findings to enhance worker skills and work engagement in the future. A critical consideration in this process will be ensuring that workers perceive such assessments as fair and acceptable. Furthermore, we intend to investigate the impact on workers' behavior when incentive structures are adjusted based on their skill levels.

7. CONCLUSIONS

In this study, we focused on the challenge of requesters' workload when evaluating workers in open-ended tasks by comparing two different methods: a peer-based method and an automated method. In our experiment evaluating worker-generated image captions, evaluations via the automated method using LLMs achieved a closer alignment with the requesters' evaluations than those via the peer-based method in terms of correctness and detail of captions. Conversely, human evaluators demonstrated a superior capability in capturing subtle nuances within the text. The findings of this research are expected to be utilized to enhance workers' skills and motivation in the future.

REFERENCES

[1] Kulkarni, A., Can, M., & Hartmann, B. (2012) "Collaboratively crowdsourcing workflows with turkomatic", *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*, pp.1003-1012.

[2] Hara, K., Le, V., & Froehlich, J. (2013) "Combining crowdsourcing and google street view to identify street-level accessibility problems", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pp.631-640.

[3] Chai, L., Sun, H., & Wang, Z. (2022) "An error consistency based approach to answer aggregation in open-ended crowdsourcing", *Information Sciences*, Vol. 608, pp.1029-1044.

[4] Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012) "Shepherding the crowd yields better work", *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*, pp.1013-1022.

[5] Saha Bhattacharya, B., Mandal, B., Biswas, A., & Bhattacharyya, M. (2022) "Improving Character Recognition by the Crowd Workers via Corrective Feedback", *2022 IEEE International Conference on Big Data (Big Data)*, pp.3982-3985.

[6] Tang, W., Yin, M., & Ho, C.-J. (2019) "Leveraging Peer Communication to Enhance Crowdsourcing", *The World Wide Web Conference (WWW '19)*, pp.1794-1805.

[7] Chiang, C.-W., Kasunic, A., & Savage, S. (2018) "Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth", *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2, No. CSCW, Article 37, pp.1-17.

[8] Nicol, D., Thomson, A., & Breslin, C. (2014) "Rethinking feedback practices in higher education: a peer review perspective", *Assessment & Evaluation in Higher Education*, Vol. 39, No. 1, pp.102-122.

[9] Zhu, H., Dow, S. P., Kraut, R. E., & Kittur, A. (2014) "Reviewing versus doing: learning and performance in crowd assessment", *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*, pp.1445-1455.

[10] de Alfaro, L. & Shavlovsky, M. (2014) "CrowdGrader: a tool for crowdsourcing the evaluation of homework assignments", *Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE '14)*, pp.415-420.

[11] Whiting, M. E., Gamage, D., Gaikwad, S. S., Gilbee, A., Goyal, S., Ballav, A., Majeti, D., Chhibber, N., Richmond-Fuller, A., Vargus, F., Sarma, T. S., Chandrakanthan, V., Moura, T., Salih, M. H., Bayomi Tinoco Kalejaiye, G., Ginzberg, A., Mullings, C. A., Dayan, Y., Milland, K., Orefice, H., Regino, J., Parsi, S., Mainali, K., Sehgal, V., Matin, S., Sinha, A., Vaish, R., & Bernstein, M. S. (2017) "Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms", *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, pp.1902-1913.

[12] Yang, T. & Miceli, M. (2024) "'Guilds' as Worker Empowerment and Control in a Chinese Data Work Platform", *Proceedings of the ACM on Human-Computer Interaction*, Vol. 8, No. CSCW2, ACM, Article 365, pp.1-27.

[13] Itano, R., Tanitsu, H., Bamba, M., Noseyama, R., Kohiga, A., & Koita, T. (2026) "Hierarchical Worker Evaluation Based on Requester's Subjective Criteria in Open-Ended Crowdsourcing Tasks", *International Journal on Cybernetics & Informatics (IJCI)*, Vol.15, No.1, pp. 1-12.

[14] Aguirre, C. A., Mahmood, A., & Huang, C.-M. (2022) "Crowdsourcing Thumbnail Captions via Time-Constrained Methods", *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*, pp.36-48.

[15] Chan, A., Di, C., Rupertus, J., Smith, G. D., Nagaraj Rao, V., Horta Ribeiro, M., & Monroy-Hernández, A. (2025) "Redefining Research Crowdsourcing: Incorporating Human Feedback with LLM-Powered Digital Twins", *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, Article 454, pp.1-10.

[16] Jones, J., Mo, L., Fosler-Lussier, E., & Sun, H. (2024) "A Multi-Aspect Framework for Counter Narrative Evaluation using Large Language Models", *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp.147-168.

[17] Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., & Sun, L. (2024) "MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark", *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Article 254, pp.1-34.

[18] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023) "Judging LLM-as-a-judge with MT-bench and Chatbot Arena", *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Article 2020, pp.1-29.

APPENDIX

Table A. Individual Worker Scores across Three Criteria and Evaluation Methods

Worker #	Fluency			Correctness			Detail		
	Req	Peer	LLM	Req	Peer	LLM	Req	Peer	LLM
1	8.50	8.83	9.03	8.50	8.58	7.93	9.50	8.50	7.40
2	8.33	7.33	8.63	9.00	6.75	7.03	9.00	6.75	7.00
3	8.17	10.00	9.43	8.67	9.92	8.53	7.50	9.25	7.23
4	7.83	9.83	9.03	8.50	9.08	7.73	6.50	8.33	6.80
5	7.17	7.83	8.07	6.00	7.25	7.13	6.00	7.33	5.17
6	7.50	8.67	8.37	5.33	8.58	6.43	6.33	7.92	5.13
7	7.00	8.58	9.03	3.33	5.50	4.10	6.00	4.17	2.93
8	6.33	8.42	8.67	3.83	6.67	5.73	4.00	3.92	3.53
9	6.33	8.08	8.77	4.33	5.75	5.33	3.00	4.08	3.30
10	5.50	5.42	7.70	5.83	6.08	5.90	1.67	2.17	2.27
11	5.00	6.42	6.67	4.67	7.42	5.40	1.33	4.08	2.27
12	2.50	5.83	6.20	2.67	5.33	4.93	0.67	2.67	1.67
13	2.17	4.00	3.83	2.17	4.50	3.07	1.00	2.08	1.00
14	1.33	3.50	2.60	1.17	4.25	3.50	0.83	2.00	1.37

Note: Req = Requester Scores, Peer = Peer Scores, LLM = LLM Scores