Simulation-based Exploration for Aggregation Algorithms in Human+AI Crowd: What factors should we consider for better results?

Takumi Tamura[1], **Hiroyoshi Ito**[1], **Satoshi Oyama**[2], **Atsuyuki Morishima**[1]

[1]University of Tsukuba, Japan
[2]Nagoya City University, Japan
tamura.takumi.ap@alumni.tsukuba.ac.jp, ito@slis.tsukuba.ac.jp, oyama@ds.nagoya-cu.ac.jp, mori@slis.tsukuba.ac.jp

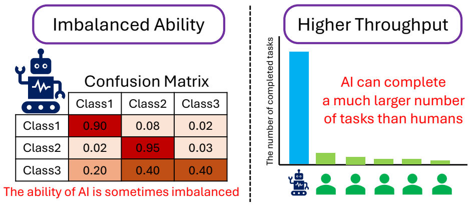
Abstract

With recent advances in AI technology, such as Large Language Models, the idea of human and AI workers performing crowdsourcing tasks together is actually being considered in some papers. However, it is still unclear what the optimal algorithm is for aggregating their responses. Interestingly, previous works suggest that the optimal algorithms are different between the human-only and human+AI crowd situations. We explore the factors influencing the aggregation process in the human+AI crowd and assume two prominent differences between humans and AI workers: (1) the ability of AI workers is often extremely imbalanced, and (2) AI workers can complete a much larger number of tasks than humans. Given the many factors that influence aggregation results, there are limitations to evaluating them using real-world datasets. This paper attempts to explore critical considerations for the human+AI crowd aggregation using a simulation-based approach.

Introduction

With the rapid growth of AI technologies such as Large Language Models (LLMs), replacing human workers with AI in annotation tasks is the focus of attention (Alizadeh et al. 2023; He et al. 2023; Zhu et al. 2023). The facts suggest the importance of the idea that AIs participate in crowdsourcing together with human workers as "AI workers" (Amer-Yahia et al. 2020). Kobayashi et al. defined the situation where we have not only humans but also black-box AI agents whose abilities are unknown, both of which complete parts of a given set of tasks as the "Human+AI Crowd" (Kobayashi, Wakabayashi, and Morishima 2021).

This paper addresses how aggregation algorithms behave in the human+AI crowd situation where duplicate classification tasks are assigned to human and black-box AI workers, and their responses are aggregated for higher-quality results. Several experimental studies have attempted to aggregate human and AI responses (Li 2024; He et al. 2024). They added one LLM worker into the human worker pool and aggregated their results by existing algorithms such as the Dawid–Skene (DS) model (Dawid and Skene 1979), the OneCoin model (Zhang et al. 2014), and GLAD (Whitehill et al. 2009). Their results suggested the OneCoin model or GLAD is better than the DS model. Interestingly, the results conflicted with the other previous study in a human-only crowd: Zheng et al. have concluded the DS model was better than those (Zheng et al. 2017).



	Class1	Class2	Class3
Class1	0.90	0.08	0.02
Class2	0.02	0.95	0.03
Class3	0.20	0.40	0.40

Figure 1: Two factors prominent in AI workers that can affect the results of aggregation algorithms

However, those studies did not unveil the factors that cause such conflicts because there are many factors that influence aggregation results, and experiments using real-world datasets have difficulty covering an extensive set of those crowdsourcing settings. For example, (1) the number of task assignments to each human worker, (2) the number of duplicates for aggregation, and (3) the number of classes in the multi-classification tasks.

In addition to those factors, the differences between human and AI workers should also be considered. We assume the following two prominent factors affect the aggregation quality (Figure 1).

(1) Imbalanced ability. It is well known that AIs trained on imbalanced data tend to produce imbalanced predictions (Branco, Torgo, and Ribeiro 2016). Although some AIs have generic abilities, such as LLMs, some papers report that their performance worsens in specific domains (Lin, Hilton, and Evans 2022). These imbalances result in each AI worker having a significantly imbalanced confusion matrix, which differs not only from human workers but also from other AI workers.

(2) Higher throughput. AI workers can complete a much larger number of tasks than humans. In contrast, human workers generally complete a small number of tasks, as suggested by previous investigations focusing on Amazon Mechanical Turk workers (Hara et al. 2018). Consequently, there is a significant gap between human and AI workers in the number of tasks completed.

Research Questions. This paper intends to move the dis-

Presented at the Works-in-Progress and Demonstrations track, AAAI HCOMP 2024. Copyright by the author(s).

cussion one step forward with an extensive set of simulation experiments instead of real-world experiments. Our research questions are the following: **(RQ1)** What important factors should we consider when aggregating the human and AI responses to obtain better results? **(RQ2)** What algorithms should we use considering the influences of those important factors?

Contributions and Key Findings. Our results clearly show that (1) the DS model is better in many cases when AI has imbalanced ability but the OneCoin model and GLAD are better than the DS model in some other cases (Figure 2), which explains why experimental results in the previous studies showed the OneCoin model or GLAD was better; the case in which the DS model does not perform well is when the number of tasks assigned to human workers is small. This is because the DS model tends to underestimate human workers who only work on a small number of tasks. (2) A hybrid algorithm of the DS and OneCoin model to address this issue can mitigate the negative effect of the DS model so that it performs better in most cases (Figure 3).

Limitations and Future Work. This paper is work-in-progress, and there are several limitations. Our simulation results are compatible with the settings of only two published papers on real-world crowdsourcing with AI workers. We need to compare our results with more real-world crowdsourcing results. In addition, to simplify the factors affecting the experimental results, we assumed that only one type of AI worker joins human workers because the interaction of multiple types of AI workers would be extremely complex.

Supplemental Materials. The technical appendix and Python implementations are available in https://github.com/crowd4u/hcomp24-wip-tamura.

Related Work

Aggregations in the Human+AI Crowd. Research in this category assumes that the behavior of AI workers is not known in advance (Kobayashi, Wakabayashi, and Morishima 2021). He et al. compared several aggregation algorithms when adding responses from GPT-4 to the human responses (He et al. 2024), and Li also conducted similar research (Li 2024). These two studies were based on the reports that LLMs match the performance of human crowd workers (Allen, He, and Gadiraju 2023; Alizadeh et al. 2023; He et al. 2023; Zhu et al. 2023). However, they considered only limited conditions with available LLM-based AI services at that time because it is difficult to apply their approaches to study under an extensive set of configurations in crowdsourcing, such as the number of tasks per human worker and the imbalanced ability of AI. Kobayashi et al. (Kobayashi, Wakabayashi, and Morishima 2021) took a more general assumption that AI workers can be any black-box AI workers other than LLMs, but they only dealt with task assignments and did not discuss aggregation.

Aggregation Algorithms. Besides standard models such as the DS (Dawid and Skene 1979), OneCoin (Zhang et al. 2014), and GLAD (Whitehill et al. 2009), there are other aggregation methods that have been extended to address the purposes related to our paper. Examples consider the difficulty of estimating confusion matrices and correct for them using a common confusion matrix for all or some of the workers (Liu and Wang 2012; Imamura, Sato, and Sugiyama 2018), and mitigate bias in human responses toward certain classes, seen in such as sentiment analysis (Wu et al. 2023; Zhang et al. 2017). Our previous work extended the DS model to utilize the uncertainty of AI workers (Tamura et al. 2024). However, they have not unveiled the key factors when aggregating the responses from human and AI workers, so none of the above give answers to our research questions.

Settings		Options (One is chosen)
#Tasks		3,000
#Class	c	2,4,8
#Tasks per Human Workers	t	5,10,20,30,50,100
#Human Task Duplicates	r	3,5,10
#Tasks per AI workers	g	5,10,20,30,50,100, 200, 300,500,1000,1500,3000 (Where $g \geq t$)
Type of AI Workers' Ability		balanced or imbalanced
#Total Cases		**1,026 cases**

Table 1: The options of our simulation model

Experimental Setup

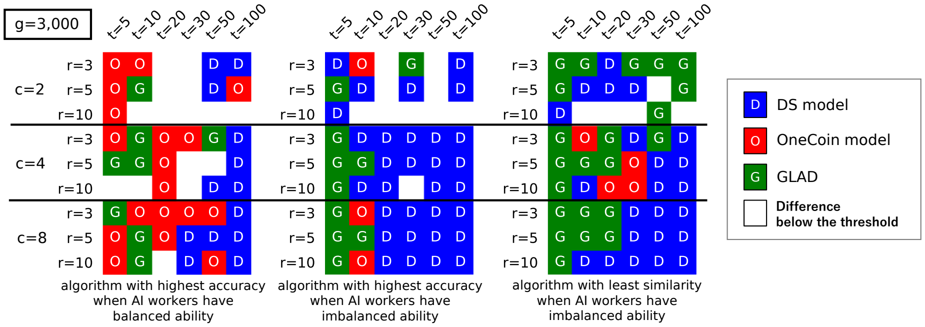
We developed the simulation model for aggregation algorithms in the Human+AI crowd. The idea behind this model is that it has a task-feature-space-centered design so that (1) workers have realistic sets of confusion matrices and (2) we can implement complex behaviors of AI workers, such as imbalanced ability, in a transparent way. These aspects are essential for simulating aggregation algorithms in the human+AI crowd.

The simulation model generates a set of multiple classification tasks based on a synthesized dataset in the 2-dimensional feature space and yields the results of simulated workers' completing the tasks. The human and AI workers are generated by representing their ability differences.

Table 1 shows the settings for the simulation model. We created 3 types of multi-class classification tasks and the model allowed us to set t, r, and g. Furthermore, we made two types of AI workers (with balanced or imbalanced ability). AI workers with imbalanced ability have the imbalanced confusion matrix like Figure 1, in other words, they have problems classifying the tasks belonging to the specific classes. There are a total of 1,026 cases of setting. The detailed explanation, parameters, and Python implementation of the model are available in the supplemental materials.

Experimental Results and Discussion

Experiment 1. We first conducted an experiment to address **RQ1**. We compared three aggregation algorithms in Crowd-Kit (Ustalov, Pavlichenko, and Tseitlin 2024) with a crowd consisting of human workers and AI workers in the simulation model. We added one type of AI worker (with balanced or imbalanced ability) to the worker set and aggregated their responses. For each type of AI worker, we conducted simulations with combinations of the parameters c, t, r, g. The


g=3,000
D DS model
O OneCoin model
G GLAD
Difference below the threshold

g=3,000		t=5	t=10	t=20	t=30	t=50	t=100
c=2	r=3	O	O			D	D
	r=5	O	G			D	O
	r=10	O					
c=4	r=3	O	G	O	O	G	D
	r=5	G	G	O			D
	r=10			O		D	D
c=8	r=3	G	O	O	O	O	D
	r=5	O	G	O	D	D	D
	r=10	O	G		D	O	D

algorithm with highest accuracy when AI workers have balanced ability

		t=5	t=10	t=20	t=30	t=50	t=100
c=2	r=3	D	O		G		D
	r=5	G	D		D		D
	r=10	D					
c=4	r=3	G	D	D	D	D	D
	r=5	G	G	D	D	D	D
	r=10	G	D	D		D	D
c=8	r=3	G	O	D	D	D	D
	r=5	G	G	D	D	D	D
	r=10	G	O	D	D	D	D

algorithm with highest accuracy when AI workers have imbalanced ability

		t=5	t=10	t=20	t=30	t=50	t=100
c=2	r=3	G	G	D	G	G	G
	r=5	G	D	D	D		G
	r=10	D				G	
c=4	r=3	G	O	G	D	G	D
	r=5	G	G	G	O	D	D
	r=10	G	D	O	O	D	D
c=8	r=3	G	G	G	D	D	D
	r=5	G	G	G	D	D	D
	r=10	G	D	D	D	D	D

algorithm with least similarity when AI workers have imbalanced ability

Figure 2: The aggregation algorithm with the highest accuracy or least similarity for each setting. The similarity (Cohen's κ) between the aggregation result and the AI's answer of the tasks belonging to the classes that AI workers have problems classifying is measured when the AI workers have imbalanced ability. Note that the cases where the difference between maximum and minimum value is below the threshold (< 0.001) are shown as blanks.

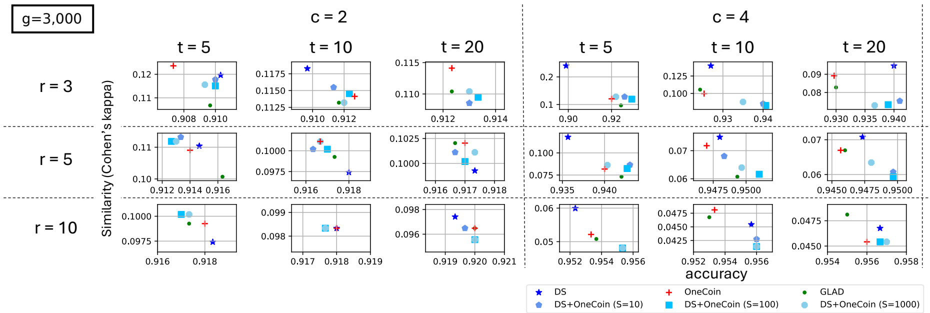


Figure 3: Accuracy and similarity about AI workers with imbalanced ability in Experiment 2 when t is small and $g = 3,000$. The X-axis shows accuracy. The Y-axis shows similarity (Cohen's κ). The algorithm in the lower right corner is better.

accuracy of aggregation results was evaluated. In addition, the similarity (Cohen's κ) between the aggregation results and the responses from AI workers with imbalanced ability was also evaluated. It was measured in the tasks belonging to the specific classes that they have problems classifying.

As the larger g is common in the Human+AI crowd (like Figure 1), we focused on the case when $g = 3,000$ (maximum value)[1]. Figure 2 shows the algorithm with the highest accuracy or least similarity. It describes that (1) The DS model was better when AI workers have imbalanced ability, but (2) the DS model outputs lower-quality results when t is small. Even in the case of imbalanced ability, the DS model performed worse when t was small, and the results were similar to the imbalanced predictions of the AI worker.

Overall, the t and the imbalanced ability of AI workers are key factors in the human+AI crowd. The DS model is better considering the imbalanced ability of AI workers, but small t provides negative effects on its results. Therefore, we need to mitigate the negative effects of the DS model with small t. We hypothesized that the negative effects arose because the DS model could not accurately estimate the confusion matrices of the human workers in those cases; they were $c \times c$ matrices (where c was the number of classes), and maximum likelihood estimation of many latent variables with few samples is generally difficult (Note that the OneCoin model and GLAD estimate workers' ability as a single parameter).

Experiment 2. This section addresses **RQ2**. The results of Experiment 1 suggest that the DS model would perform better in the human+AI crowd, *if* it could improve the estimation of confusion matrices when t is small. To verify the hypothesis, we conducted additional experiments with the **DS+OneCoin model**[2], a hybrid algorithm of the DS and OneCoin model that attempts to estimate confusion matrices well when t is small. The experiment settings were the same as those for Experiment 1 (shown in Table 1).

Figure 3 shows a part of experimental results[1] about AI workers with imbalanced ability. It shows that the DS+OneCoin model (the parameter S adjusts the weights of the DS and OneCoin model) performed well when t is small. These results support the hypothesis and give us deep insights to explore the optimal aggregation algorithm in the Human+AI crowd.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number JP21H03552, JP22H00508, JP23H03405, JST CREST Grant Number JPMJCR21D1, and JPMJCR22M2.

[1] All of the experimental results including other cases are available in the supplemental materials.

[2] The detailed formalization and Python implementation are available in the supplemental materials.

References

Alizadeh, M.; Kubli, M.; Samei, Z.; Dehghani, S.; Bermeo, J. D.; Korobeynikova, M.; and Gilardi, F. 2023. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. *arXiv preprint arXiv:2307.02179*.

Allen, G.; He, G.; and Gadiraju, U. 2023. Power-up! What Can Generative Models Do for Human Computation Workflows? *arXiv preprint arXiv:2307.02243*.

Amer-Yahia, S.; Basu Roy, S.; Chen, L.; Morishima, A.; Abello Monedero, J.; Bourhis, P.; Charoy, F.; Danilevsky, M.; Das, G.; Demartini, G.; Elbassuoni, S.; Gross-Amblard, D.; Hoareau, E.; Inoguchi, M.; Kenworthy, J.; Kitahara, I.; Lee, D.; Li, Y.; Borromeo, R. M.; Papotti, P.; Rao, R.; Roy, S.; Senellart, P.; Tajima, K.; Thirumuruganathan, S.; Tommasi, M.; Umemoto, K.; Wiggins, A.; and Yoshida, K. 2020. Making AI Machines Work for Humans in FoW. *ACM SIGMOD Record*, 49(2): 30–35.

Branco, P.; Torgo, L.; and Ribeiro, R. P. 2016. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.*, 49(2).

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 20–28.

Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, C.; and Bigham, J. P. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

He, X.; Lin, Z.; Gong, Y.; Jin, A.-L.; Zhang, H.; Lin, C.; Jiao, J.; Yiu, S. M.; Duan, N.; and Chen, W. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. *arXiv preprint arXiv:2303.16854*.

He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Imamura, H.; Sato, I.; and Sugiyama, M. 2018. Analysis of Minimax Error Rate for Crowdsourcing and Its Application to Worker Clustering Model. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, 2147–2156.

Kobayashi, M.; Wakabayashi, K.; and Morishima, A. 2021. Human+AI Crowd Task Assignment Considering Result Quality Requirements. In *Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 9, 97–107.

Li, J. 2024. A Comparative Study on Annotation Quality of Crowdsourcing and LLM Via Label Aggregation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6525–6529.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, 3214–3252.

Liu, C.; and Wang, Y.-M. 2012. TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, 17–24.

Tamura, T.; Ito, H.; Oyama, S.; and Morishima, A. 2024. Influence of AI's Uncertainty in the Dawid-Skene Aggregation for Human-AI Crowdsourcing. In *Proceedings of the 19th International Conference on Information (iConference)*, volume 3, 232–247.

Ustalov, D.; Pavlichenko, N.; and Tseitlin, B. 2024. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96): 6227.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*, volume 22, 2035–2043.

Wu, M.; Li, Q.; Yang, F.; Zhang, J.; Sheng, V. S.; and Hou, J. 2023. Learning from biased crowdsourced labeling with deep clustering. *Expert Systems with Applications*, 211.

Zhang, J.; Sheng, V. S.; Li, Q.; Wu, J.; and Wu, X. 2017. Consensus algorithms for biased labeling in crowdsourcing. *Information Sciences*, 382–383: 254–273.

Zhang, Y.; Chen, X.; Zhou, D.; and Jordan, M. I. 2014. Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 1, 1260–1268.

Zheng, Y.; Li, G.; Li, Y.; Shan, C.; and Cheng, R. 2017. Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment*, 10(5): 541–552.

Zhu, Y.; Zhang, P.; Haq, E.-U.; Hui, P.; and Tyson, G. 2023. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. *arXiv preprint arXiv:2304.10145*.