A Brief Summary of the Study "If in a Crowdsourced Data Annotation Pipeline, a GPT-4"

Zeyu He
The Pennsylvania State University
University Park, PA, USA
zmh5268@psu.edu

Chieh-Yang Huang
The Pennsylvania State University
University Park, PA, USA
chiehyang@alumni.psu.edu

Chien-Kuang Cornelia Ding
University of California, San Francisco
San Francisco, CA, USA
cornelia.ding@ucsf.edu

Shaurya Rohatgi
The Pennsylvania State University
University Park, PA, USA
szr207@psu.edu

Ting-Hao 'Kenneth' Huang
The Pennsylvania State University
University Park, PA, USA
txh710@psu.edu

ABSTRACT

Recent studies indicated GPT-4 outperforms online crowd workers in data labeling accuracy, notably workers from Amazon Mechanical Turk (MTurk). However, these studies were criticized for deviating from standard crowdsourcing practices and emphasizing individual workers' performances over the whole data-annotation process. This paper compared GPT-4 and an ethical and well-executed MTurk pipeline, with 415 workers labeling 3,177 sentence segments from 200 scholarly articles using the CODA-19 scheme. Two worker interfaces yielded 127,080 labels, which were then used to infer the final labels through eight label-aggregation algorithms. Our evaluation showed that despite best practices, MTurk pipeline's highest accuracy was 81.5%, whereas GPT-4 achieved 83.6%. Interestingly, when combining GPT-4's labels with crowd labels collected via an advanced worker interface for aggregation, 2 out of the 8 algorithms achieved an even higher accuracy (87.5%, 87.0%). Further analysis suggested that, when the crowd's and GPT-4's labeling strengths are complementary, aggregating them could increase labeling accuracy.[1]

CCS CONCEPTS

• **Information systems → Crowdsourcing**; • **Applied computing → Annotation**; • **Computing methodologies → Natural language processing**; • **Human-centered computing → Human computer interaction (HCI)**.

ACM Reference Format:

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao 'Kenneth' Huang. 2024. A Brief Summary of the Study "If in a Crowdsourced Data Annotation Pipeline, a GPT-4". In *Proceedings of LLMs as Research Tools: Applications and Evaluations in HCI Data Work, a workshop at CHI 2024 (LART@CHI '24).* ACM, New York, NY, USA, 11 pages.

[1]This document is a short summary of the study "If in a Crowdsourced Data Annotation Pipeline, a GPT-4" [5], prepared and submitted by its authors to the "LLMs as Research Tools: Applications and Evaluations in HCI Data Work" workshop at CHI 2024.

LART@CHI '24, May 12, 2024, Honolulu, HI, USA

1 INTRODUCTION

This paper presents a **holistic investigation that compares the data labeling abilities of GPT-4 and a well-executed, ethical MTurk data-annotation process.** We had 415 Mturk workers label 3,177 sentence segments in 200 scholarly paper published in 2022 or later, where each segment collected 20 labels. We used the CODA-19 dataset's 5-class label scheme [7], categorizing sentence segments by their research aspects: Background, Purpose, Method, Finding/Contribution, and Other. We further experimented with two distinctly designed worker interfaces, recognizing the potential biases of any particular annotation interface [9, 11]. A total of 127,080 labels was gathered in the study.[2] We then applied 8 label aggregation algorithms to determine the final labels and compared the labeling accuracies with GPT-4. We found that, even with the best crowdsourcing practices, **MTurk's top-performing pipeline' accuracy of 81.5% did not surpass GPT-4's 83.6% ($p<0.05$).** Interestingly, **when combining GPT-4's labels with crowd labels collected via an advanced worker interface for aggregation, 2 out of the 8 algorithms achieved a higher accuracy (87.5% with $p<0.01$ and 87.0% with $p<0.01$) compared to GPT-4's standalone performance (83.6%)**. Further analysis suggested that, when the crowd's and GPT-4's labeling strengths are complementary– with crowds better at labeling the Finding/Contribution class and GPT-4 excelling in all other classes– aggregating them could further increase labeling accuracy.

The contribution of this paper is three-fold. First, it responds to recent speculations about GPT-4's labeling ability surpassing online crowd workers by focusing on the performance of holistic crowdsourcing pipelines, an area previously overlooked. Second, our study highlights the value of crowdsourced labels in scenarios where GPT-4's accuracy generally outperforms yet complements crowd efforts, demonstrating that adding crowd labels can further enhance accuracy. Third, this study sheds light on the evolving role and best practices for crowdsourcing in the era of Large Language Models (LLMs), particularly when LLMs often exhibit superior labeling accuracy compared to crowd workers.

[2]We will keep the dataset offline for a minimum of one year to avoid data contamination concerns of large language models pre-trained using data on the public Internet.Access to the dataset will be email-exclusive, and we will mandate recipients not to upload it to the Internet. Before sharing the data, we will hash the worker IDs to ensure they are unique within the dataset but unrecognizable as real worker IDs.

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI MV	.713	.281	.403	.149	.525	.232	.368	.599	.456	.772	.507	.612	1.000	.286	.444	.477	.285
Advanced UI MV	.598	.307	.405	.112	.438	.179	.373	.634	.469	.815	.425	.559	-	0	-	.442	.259
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764
GPT-4 (t=1.0)	.859	.903	.881	.493	.829	.619	.766	.876	.818	.978	.783	.870	.346	.857	.493	.833	.760
CS Expert	.900	.801	.848	.541	.853	.662	.856	.801	.828	.913	.915	.914	1.000	.619	.765	**.859**	.788

Table 1: Performance using Bio Expert as the gold standard. The CS Expert achieves the highest accuracy of 85.9% over all models. GPT-4 at temperature of .2 and 1.0 have accuracies of 83.6% and 83.3%. Basic and Advanced Majority Vote had accuracy of 47.7% and 44.2%.

2 METHODOLOGY

In this paper, we followed the bet practice of crowdsourcintg to use MTurk to label new data, applied a variety of label aggregation techniques to induce final labels, and compared the result with GPT-4. This section detailed the procedure. Previous research indicates worker interface design can influence performance [9, 11]; thus, we tested two different interfaces in our study.

2.1 Annotation Scheme and Data

Annotation Scheme and Instruction. We aimed to compare MTurk and GPT-4's ability to label text items, as GPT-4 currently performs best with text rather than with video or images. For our study, we chose the CODA-19 label scheme [7], which categorizes sentence segments in paper abstracts into research aspects, *i.e.*, Background, Purpose, Method, Finding/Contribution, and Other. We obtained the detailed annotation instructions via CODA-19's GitHub repository[3] and used it in our study.

This task was picked for its balanced difficulty: It demands reading scholarly articles, making it more challenging than basic sentiment labeling. However, it is not as hard as expert-only labeling tasks like disease mentions, and MTurk workers have successfully completed it before [2, 7].

Data. The original CODA-19 dataset [7] contains biomedical papers published before April 2020 extracted from the COVID-19 Open Research Dataset (CORD-19) dataset [15]. In this study, we sampled papers from the most recent release of the CORD-19 dataset, dated June 2, 2022, which housed around one million documentations. To prevent our test data from overlapping with OpenAI GPT's training data, we limited our study to documents published after ChatGPT's last knowledge update in September 2021, focusing on 2022 publications or later. We used textttlangdetect[4] to identify and retain 123,881 English papers in our dataset.

For our main study, we randomly sampled 200 papers from this dataset as the test set. We segmented the abstracts of these papers into 3,177 sentence segments, averaging 15.89 segments per abstract, following CODA-19's approach [7].

For developing worker interfaces (Section 2.2), we also randomly sampled 200 different papers from the dataset as the interface development set.

[3]CODA-19: COVID-19 Research Aspect Dataset: https://github.com/windx0303/CODA-19

[4]Language detection library in Python: https://github.com/fedelopez77/langdetect

2.2 Collecting Labels via Amazon Mechanical Turk

Worker Interfaces. Prior studies suggested that the design of the worker interface would impact annotation performances on MTurk [9, 11]. To address potential biases, we tested two interfaces in our study. Both displayed the original CODA-19 instructions but were independently designed by different individuals:

- **Basic Worker Interface (Figure 2):** An author of this paper, unfamiliar with designing interfaces for MTurk tasks, was tasked with creating a worker interface using the original CODA-19 annotation instructions, including examples and FAQs. We emphasized simplicity and usability in the design.
- **Advanced Worker Interface (Figure 3):** It is the original interface that was used for constructing the CODA-19 dataset [7], designed by a crowdsourcing expert with extensive experience in designing MTurk task interfaces.

Both interfaces show the original CODA-19 annotation instructions. We did not explicitly tell workers that they were part of an experiment comparing MTurk pipelines with GPT; we simply stated it was a data labeling task. This approach was chosen to replicate a typical data labeling scenario.

Posting Tasks in Batches and Monitoring Label Quality. We divided 200 abstracts (see Section 2.1) into four batches of 50, posting one at a time. For each abstract, we created two HITs: one with the basic interface and the other with the advanced interface. We recruited 20 workers via 20 assignments (from the qualified pool of 400) for each HIT. Once a batch was completed, we assessed label quality and removed qualifications from underperforming workers to prevent them from accessing our future HITs. The "CS Expert" manually labeled only 10 abstracts per batch. We used these labels to compute three worker quality control statistics: *(i)* label accuracy, based on only 10 manually labeled abstracts per batch, *(ii)* probability of agreeing with the majority label, and *(iii)* probability of labeling "Other," a rare label. For *(i)* and *(ii)*, we reviewed the bottom 30 workers' labels, and for *(iii)*, the top 30's. If we observed a worker consistently providing incorrect labels or seemingly spamming our task, we revoked their qualification.

2.3 Collecting Gold-Standard Labels Using Experts

Similar to CODA-19 [7], we worked with two experts, a biomedical expert (Bio Expert), and a computer science expert (CS Expert). Both of these experts, who are also co-authors of this paper, manually annotated the entire test set of 200 abstracts from our MTurk study using the advanced interface. The inter-annotator agreement (Cohen's kappa) between the two was 0.788.

The **"Bio Expert"**in Table 1 is Dr. Chien-Kuang Cornelia Ding. She is a faculty member in the Department of Pathology at the University of California, San Francisco. Dr. Ding possesses an M.D. and a Ph.D. in Genetics and Genomics.

The **"CS Expert"** in Table 1, the first author of the paper, is a Ph.D. student in Informatics and well-acquainted with our annotation scheme. We used a subset of the CS Expert's labels to remove underperforming workers in the annotation process (Section 2.2), replicating scenarios where gold-standard labels (Bio Expert's) are not fully available (yet), and requesters must label part of the data themselves.

2.4 Annotating Data Using GPT-4

We used the full worker instruction from the original CODA-19 dataset as GPT-4's prompt for our data labeling [7]. Our initial perception was that GPT-4 underperformed in this specific task, given that it was reported to have inferior performance compared to the SciBERT model fine-tuned on the CODA-19 dataset [3]. However, we noticed that the prompt used in the said study did not contain the entire abstract [3], which might have led GPT-4 to rely on partial context for predictions. So, we modified the prompt to include the full abstract for a zero-shot approach (Table 10). Following prior studies that compared GPT-4's zero-shot capabilities with crowd workers [12], we tested GPT-4 using both high (1.0) and low (0.2) temperature settings.

2.5 Label Cleaning Strategies

Label Cleaning Strategies. As described in Section 2.2, we removed underperforming workers' qualifications after each data batch so they can not participate in future batches. We explored three strategies in this paper:

- **All:** Retain every collected label without any exclusions.
- **Exclude-By-Worker:** Exclude labels from any MTurk worker who was ever removed.
- **Exclude-By-Batch:** Only exclude a label if its annotator was removed during that specific data batch. This means if a worker was removed from a given batch, we only exclude their labels from that batch but retain those from prior batches.

Only the selected labels will proceed to the follow-up label aggregation step.

2.6 Label Aggregation Methods

In our study, we explored a range of label aggregation algorithms. First, we adopted the majority voting method, including its tie-breaker approach, directly from CODA-19 [7]. Second, we utilized a series of aggregation algorithms provided by Crowd-Kit [13, 14], such as Dawid-Skene [4], One-coin Dawid-Skene [17], M-MSR (Matrix Mean-Subsequence-Reduced Algorithm) [8], Worker Agreement with Aggregate (Wawa) [1], Zero-Based Skill (ZBS) [10] and GLAD (Generative model of Labels, Abilities, and Difficulties) [16]. Finally, we also experimented with MACE (Multi-Annotator Competence Estimation) implemented by Hovy et al. [6].

3 FINDING

In this section, we first overview the comparative results of GPT-4 and MTurk pipeline with a variety of settings (Section 3.1) and then show the results of incorporating GPT-4 into MTurk pipelines (Section 3.2.)

3.1 GPT-4 vs. MTurk Pipelines

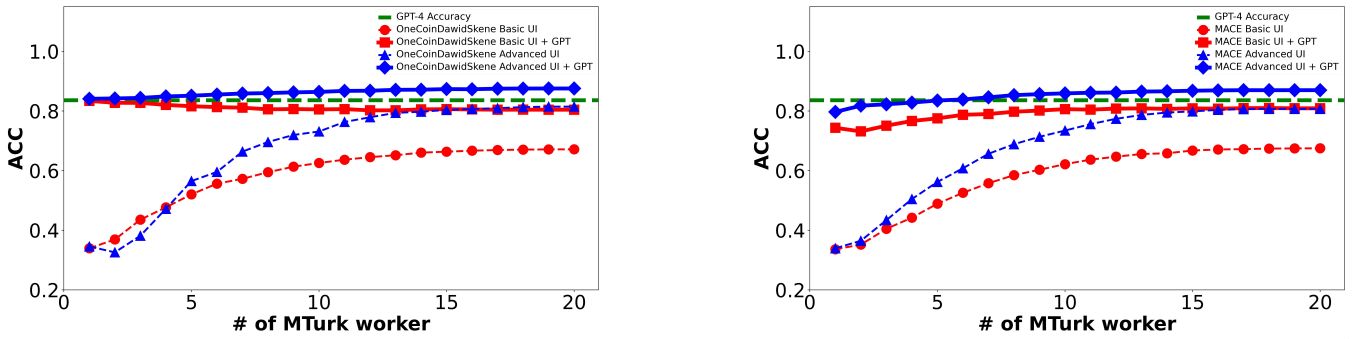
To evaluate the labeling accuracy, we used the Bio Expert's labels as the gold-standard labels. We employed the majority vote as our baseline model, as it was used in the original CODA-19 paper. Table 1 shows the results. **GPT-4 exhibited accuracies of 83.6% and 83.3% at low (0.2) and high (1.0) temperatures, respectively.** Using Majority Voting (MV), labels provided by MTurk workers from the Basic and Advanced interface groups achieved accuracies of 47.7% and 44.2%, respectively. The CS Expert achieved the highest accuracy of 85.9%.

Exclude-By-Worker is the best label-cleaning strategy. Next, we experimented with the three different label-cleaning strategies mentioned in Section 2.6. Table 2 to Table 4 show the accuracy of three strategies paired with different aggregation models. Both the Exclude-By-Worker and Exclude-By-Batch strategies enhanced the aggregation accuracy, but Exclude-By-Worker produced superior results. When pairing the One-Coin Dawid-Skene aggregation method with MTurk workers in the Advanced Interface group using the Exclude-By-Worker approach, we achieved **the best accuracy of 81.5%** (Table 3). While this surpassed other aggregation models with different strategies, it **did not exceed the performance of GPT-4 (83.6%)**. As the Exclude-By-Worker label cleaning strategy produced the best results, **we showcase results only using the Exclude-By-Worker cleaning strategy throughout the remainder of the paper.**

3.2 GPT-4 in MTurk Pipelines

Driven by a collaborative perspective on crowdwork, this paper emphasizes the importance of aggregating results from all labels for the final output. The previous section compared GPT-4 against MTurk pipelines as separate entities, and despite our efforts, nothing surpassed GPT-4's performance. It is intriguing to consider the potential impact of integrating **GPT-4 as a worker** into the label aggregation process. This subsection presents the findings.

Combining GPT-4 with crowd labels can potentially exceed GPT-4's solo performance. In our simulation study, we treated GPT-4 as an MTurk worker and selected it at t=0.2 due to its high accuracy. The results are shown in Figure 1. The x-axis in each figure indicates the number of human workers in the aggregation. A count of 5 workers, for example, refers to a mix of 5 MTurk workers and the GPT-4 model.



(a) One-Coin Dawid-Skene

(b) MACE

Figure 1: Exclude-By-Worker simulation results applied to different aggregation models. One-Coin Dawid-Skene and MACE algorithms for a combination of advanced interface results had the best accuracy and outperformed GPT-4 at temperature 0.2 (see (a) (b)).

In aggregations that included GPT-4, the Advanced Interface results using One-Coin Dawid-Skene (Figure 1a) and MACE (Figure 1b) consistently **surpassed GPT-4's performance (83.6%), peaking at an accuracy of 87.5%**. All these two settings confirm the statistical significance of the improvement, where results are detailed in Table 9. Notably, this was **the only two settings in our experiments that bested GPT-4**, demonstrating the potential and the challenge to enhance accuracy by incorporating crowd labels. This finding suggests that even a handful of crowd labels can be beneficial.

4 DISCUSSION

Crowdsourced Data Annotation Practices in the Era of LLMs. Our study demonstrates that even a meticulously crafted MTurk pipeline may not outperform the zero-shot GPT-4 in labeling accuracy. We spent weeks developing, testing, and implementing Basic and Advanced interfaces. After finalizing these interfaces, another two weeks were dedicated to posting tasks and gathering data from MTurk workers. This was mixed with significant effort to review and filter their submissions. However, even with this level of commitment, we could not match GPT-4's performance. In contrast, the efficiency of GPT-4 was outstanding. The design, testing, and execution of annotation tasks took two days. This brings us to a pivotal question: In light of the fact that LLMs can now, in some instances, outperform human annotators, **how will the practices of data annotation evolve?** While we cannot definitively answer it, we want to give a few thoughts based on our study:

- Firstly, **the value of expert-level, high-quality labels will likely rise significantly.** In our study, gold labels played a central role in several critical decisions: refining prompts for greater efficacy, choosing the most effective label-cleaning strategy, and selecting the best label-aggregation algorithms. These decisions led us to the few parameter combinations (Advanced Interface + OneCoin/MACE + incorporating GPT-4) that eventually surpassed GPT-4's performance.
- Second, the research focus might **shift from "using AI to support human labelers" to "using humans to enhance AI labeling."** Our study showed that by carefully adding a few crowd labels, GPT-4's accuracy can be improved. Given the cost and difficulty of finding expert labelers, using non-expert labels to enhance LLM's performance will likely become more critical.
- Finally, while it might appear as a nuanced point, we believe that **the Human-Computer Interaction (HCI) challenges in the human annotation process will become central again**. In our study, initial observations suggest marginal differences between the Basic and Advanced interfaces. However, the more detailed analysis show the strengths of the advanced interface. Workers using it provided more consistent labels, making it the only interface to surpass GPT-4. Given LLMs' high labeling accuracy, we will likely need even more reliable human labels in the future to boost their performance further. Developing systems that allow users, especially non-expert annotators, to perform reliably and consistently is essentially an HCI problem.

5 CONCLUSION AND FUTURE WORK

This paper evaluates GPT-4's labeling capabilities in contrast to a well-executed, ethical crowdsourcing pipeline for annotating unseen data. Utilizing the CODA-19 labeling scheme, we exhaustively tested various label-cleaning strategies, label-aggregation techniques, and interface designs on MTurk. Despite adhering to best crowdsourcing practices, the best-performing MTurk pipeline achieved an accuracy of 81.5%, slightly below GPT-4's 83.6%. Interestingly, by optimizing the combination of label aggregation techniques and interfaces, integrating GPT-4 labels with the MTurk aggregation process boosted accuracy to 87.5%.

Moving forward, our research will focus on generating a smaller set of high-quality labels via MTurk, aiming to further enhance the labeling performance of already sophisticated LLMs like GPT-4. Additionally, we will delve deeper into the influence of worker interface design on label quality to further improve LLM performance.

REFERENCES

[1] Appen. 2021. Calculating Worker Agreement with aggregate (WAWA). https://success.appen.com/hc/en-us/articles/202703205-Calculating-Worker-Agreement-with-Aggregate-Wawa-

[2] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (nov 2018), 21 pages. https://doi.org/10.1145/3274300

[3] Shreya Chandrasekhar, Chieh-Yang Huang, and Ting-Hao Huang. 2023. Good Data, Large Data, or No Data? Comparing Three Approaches in Developing Research Aspect Classifiers for Biomedical Papers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Association for Computational Linguistics, Toronto, Canada, 103–113. https://doi.org/10.18653/v1/2023.bionlp-1.8

[4] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[5] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems at CHI 2024* (Honolulu, HI, USA,) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642834

[6] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1120–1130.

[7] Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.nlpcovid19-acl.6

[8] Qianqian Ma and Alex Olshevsky. 2020. Adversarial crowdsourcing through robust rank-one matrix completion. *Advances in Neural Information Processing Systems* 33 (2020), 21841–21852.

[9] Bahareh Rahmanian and Joseph G Davis. 2014. User interface design for crowdsourcing systems. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. 405–408.

[10] Toloka. [n. d.]. ZeroBasedSkill - crowd-kit: Toloka documentation. https://toloka.ai/docs/crowd-kit/reference/crowdkit.aggregation.classification.zero_based_skill.ZeroBasedSkill/

[11] Michael Toomim, Travis Kriplean, Claus Pörtner, and James Landay. 2011. Utility of human-computer interactions: Toward a science of preference measurement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2275–2284.

[12] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588* (2023).

[13] Dmitry Ustalov, Nikita Pavlichenko, Vladimir Losev, Iulian Giliazev, and Evgeny Tulin. 2021. A general-purpose crowdsourcing computational quality control toolkit for python. In *The Ninth AAAI Conference on Human Computation and Crowdsourcing: Works-in-Progress and Demonstration Track (HCOMP 2021)*.

[14] Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2023. Learning from Crowds with Crowd-Kit. arXiv:2109.08584 [cs.HC] https://arxiv.org/abs/2109.08584

[15] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.nlpcovid19-acl.1

[16] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009).

[17] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems* 27 (2014).

A BASIC AND ADVANCED WORKER INTERFACE

In this section we show both basic and advanced user interfaces in Figure 2 and 3.

B DETAILED RESULT TABLE

In this section, we show all aggregation results for different distinct configurations tested (2 different interfaces, and 3 different cleaning strategies). Table 2 to 7 is the aggregation accuracy results on both crowd-only and crowd+gpt results for both Basic and Advanced Interfaces under different cleaning strategies. They show detailed P, R, F1, Accuracy, and Kappa for each aggregation model under different user interfaces.

C CROWD AND GPT-4 AGGREGATED RESULT TABLE WITH CONFIDENCE INTERVAL

In this section, we show the aggregation results with confidence interval for different distinct configurations tested (2 different interfaces, and 3 different cleaning strategies).

D PROMPT

Table 10 shows the zero-shot prompt we used for querying LLMs.

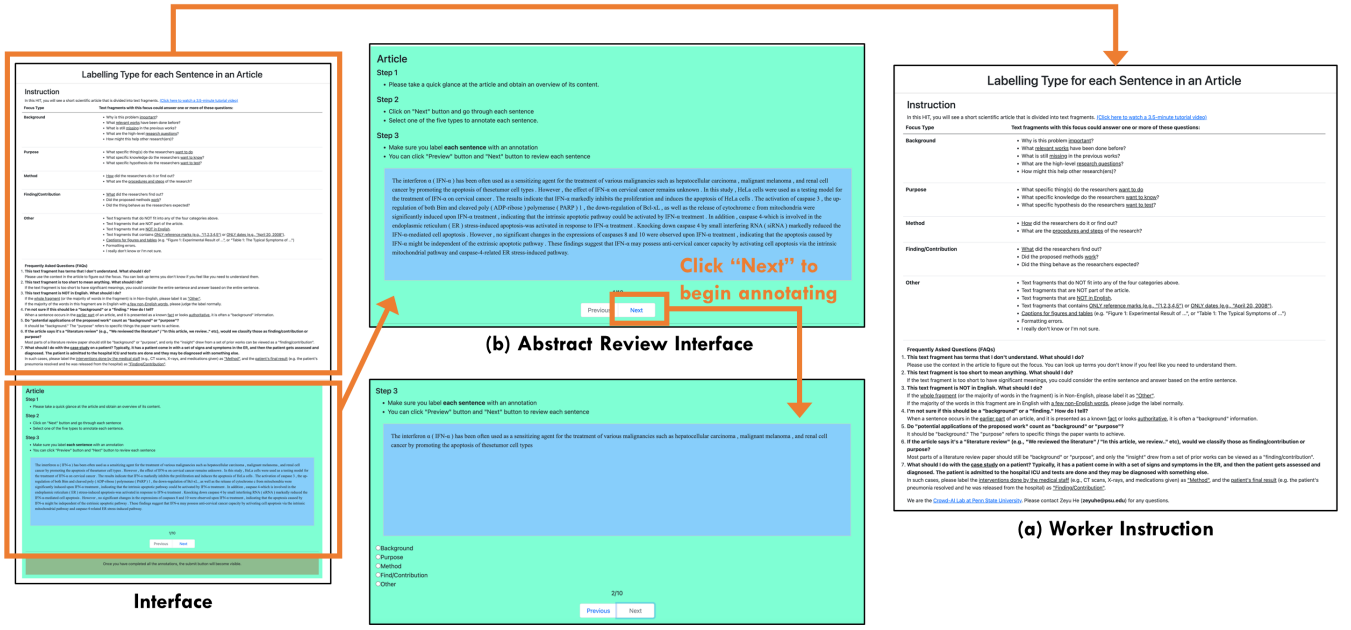


Figure 2: The basic worker interface, individually designed by one of the authors, has a focus on prioritizing task simplicity and user-friendliness.

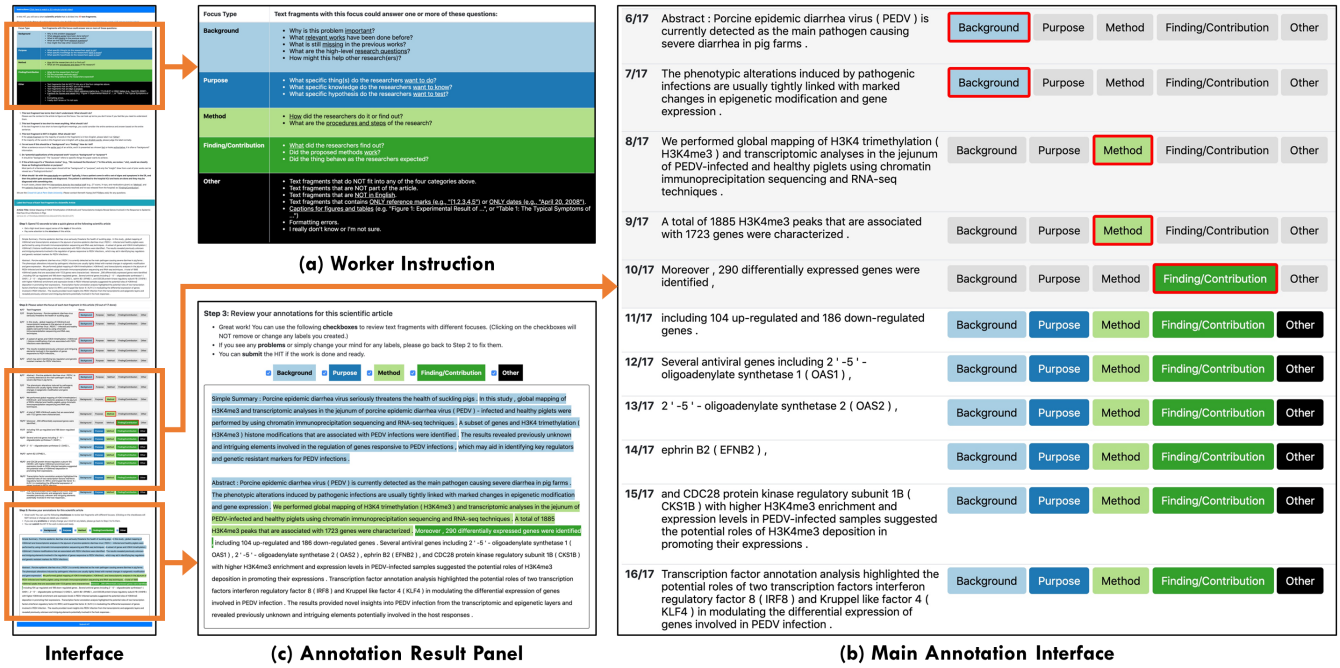


Figure 3: The advanced worker interface, adopted from CODA-19 [7], incorporates advanced features such as a visual feedback button, color-coded annotation view, and a time lock mechanism to deter hasty spam submissions.

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI																	
MV	.713	.281	.403	.149	.525	.232	.368	.599	.456	.772	.507	.612	1.000	.286	.444	.477	.285
DawidSkene	.676	.503	.577	.181	.502	.266	.504	.588	.543	.868	.561	.682	.036	.429	.066	.549	.392
OneCoin	.852	.428	.570	.328	.585	.421	.496	.737	.593	.824	.752	.787	1.000	.238	.385	.663	.504
GLAD	.767	.448	.566	.224	.645	.333	.482	.675	.563	.866	.648	.741	.360	.429	.391	.608	.451
M-MSR	.519	.337	.408	.140	.498	.218	.347	.526	.418	.768	.435	.555	.143	.238	.179	.436	.244
MACE	.733	.586	.651	.288	.631	.396	.581	.688	.630	.888	.716	.793	.165	.619	.260	**.675**	.537
Wawa	.718	.380	.497	.173	.576	.266	.408	.647	.501	.839	.536	.654	.750	.286	.414	.527	.353
ZBS	.741	.398	.518	.181	.594	.277	.426	.666	.520	.858	.559	.677	.600	.286	.387	.547	.380
Advanced UI																	
MV	.598	.307	.405	.112	.438	.179	.373	.634	.469	.815	.425	.559	-	0	-	.442	.259
DawidSkene	.565	.362	.442	.134	.475	.209	.599	.657	.627	.931	.609	.736	.046	.429	.084	.555	.401
OneCoin	.686	.404	.509	.144	.530	.226	.433	.691	.533	.881	.498	.636	-	0	-	.517	.352
GLAD	.745	.553	.635	.192	.594	.290	.554	.719	.626	.910	.637	.750	.600	.286	.387	.631	.488
M-MSR	.534	.368	.436	.113	.470	.183	.380	.587	.461	.809	.384	.521	.286	.095	.143	.428	.249
MACE	.824	.807	.815	.412	.700	.519	.757	.819	.787	.937	.803	.865	.196	.476	.278	**.798**	.707
Wawa	.586	.377	.459	.128	.516	.205	.411	.641	.501	.858	.435	.577	1.000	.095	.174	.470	.299
ZBS	.648	.438	.523	.142	.525	.224	.457	.671	.544	.883	.510	.647	1.000	.238	.385	.528	.365
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764

Table 2: All Workers Table. All models use Bio Expert as the gold standard. Baseline is the Majority Vote (MV).

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI																	
MV	.722	.342	.465	.166	.548	.254	.408	.606	.488	.794	.564	.660	.857	.286	.429	.522	.337
DawidSkene	.789	.407	.537	.228	.530	.319	.512	.682	.585	.858	.669	.752	.063	.571	.114	.604	.446
OneCoin	.817	.461	.590	.361	.599	.451	.506	.726	.597	.820	.757	.787	1.000	.238	.385	.671	.514
GLAD	.759	.483	.590	.228	.636	.336	.492	.675	.569	.861	.650	.740	.474	.429	.450	.616	.460
M-MSR	.553	.378	.449	.134	.452	.206	.345	.521	.415	.755	.438	.555	.171	.286	.214	.443	.249
MACE	.742	.589	.657	.288	.618	.392	.582	.687	.630	.881	.717	.791	.155	.619	.248	**.675**	.536
Wawa	.712	.424	.531	.183	.604	.281	.439	.647	.523	.859	.569	.684	.700	.333	.452	.555	.389
ZBS	.735	.446	.555	.202	.627	.305	.459	.674	.546	.868	.596	.707	.636	.333	.438	.580	.419
Advanced UI																	
MV	.642	.344	.448	.143	.488	.221	.390	.656	.489	.835	.490	.618	-	0	-	.490	.310
DawidSkene	.688	.411	.515	.187	.618	.288	.720	.725	.722	.936	.705	.804	.054	.476	.097	.637	.501
OneCoin	.874	.768	.818	.500	.618	.553	.703	.853	.771	.909	.853	.880	1.000	.286	.444	**.815**	.723
GLAD	.807	.610	.695	.254	.645	.365	.582	.765	.661	.927	.707	.802	.615	.381	.471	.692	.564
M-MSR	.555	.414	.474	.142	.516	.223	.388	.599	.471	.834	.433	.570	.167	.048	.074	.467	.291
MACE	.817	.822	.819	.424	.691	.525	.767	.818	.791	.939	.814	.872	.310	.619	.413	.807	.718
Wawa	.628	.460	.531	.168	.571	.260	.462	.684	.552	.890	.523	.659	1.000	.190	.320	.545	.384
ZBS	.681	.537	.600	.190	.608	.289	.518	.703	.597	.903	.579	.706	.857	.286	.429	.596	.447
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764

Table 3: Exclude-By-Worker Table. All models use Bio Expert as the gold standard. The baseline is the Majority Vote (MV). From Exclude-By-Worker results, the One-Coin Dawid-Skene aggregation model achieves the highest accuracy for both basic and advanced interfaces. Advanced One-Coin Dawid-Skene reaches 81.5% and outperforms other aggregation models in every aspect. The accuracy from advanced One-Coin Dawid-Skene almost reaches the accuracy of the GPT-4 (t=.2), 83.6%.

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI																	
MV	.688	.297	.414	.147	.535	.231	.382	.619	.472	.804	.504	.620	1.000	.286	.444	.484	.299
DawidSkene	.773	.463	.579	.197	.539	.289	.518	.616	.563	.861	.647	.739	.060	.524	.107	.592	.435
OneCoin	.839	.426	.565	.304	.594	.402	.479	.741	.582	.842	.723	.778	1.000	.238	.385	.649	.490
GLAD	.767	.447	.565	.211	.636	.317	.476	.691	.564	.871	.620	.724	.529	.429	.474	.597	.440
M-MSR	.539	.340	.417	.131	.475	.205	.314	.522	.392	.756	.392	.516	.583	.333	.424	.414	.220
MACE	.744	.580	.652	.287	.627	.394	.575	.694	.629	.885	.717	.793	.178	.619	.277	**.675**	.537
Wawa	.694	.397	.505	.166	.571	.257	.411	.643	.501	.855	.525	.651	.750	.286	.414	.524	.353
ZBS	.723	.420	.531	.178	.585	.273	.431	.663	.522	.878	.564	.687	.750	.286	.414	.553	.389
Advanced UI																	
MV	.616	.311	.413	.123	.465	.194	.376	.635	.473	.824	.451	.583	-	0	-	.458	.275
DawidSkene	.642	.354	.456	.144	.498	.223	.616	.684	.648	.933	.655	.770	.047	.429	.085	.583	.433
OneCoin	.698	.417	.522	.156	.558	.244	.449	.712	.551	.890	.517	.654	-	0	-	.536	.374
GLAD	.729	.572	.641	.202	.618	.305	.567	.726	.637	.918	.639	.753	.667	.286	.400	.639	.499
M-MSR	.481	.394	.433	.128	.493	.203	.383	.565	.456	.822	.400	.538	.143	.048	.071	.438	.258
MACE	.825	.792	.808	.404	.696	.511	.739	.807	.772	.931	.793	.856	.167	.476	.247	**.787**	.692
Wawa	.608	.383	.470	.135	.530	.215	.416	.650	.507	.866	.455	.596	1.000	.143	.250	.484	.314
ZBS	.655	.451	.534	.154	.558	.242	.468	.676	.554	.887	.525	.659	1.000	.286	.444	.542	.381
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764

Table 4: Exclude-By-Batch Table. All models use Bio Expert as the gold standard. Baseline is the Majority Vote (MV).

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI																	
MV	.798	.380	.515	.187	.604	.286	.435	.685	.532	.828	.566	.672	1.000	.286	.444	.551	.381
DawidSkene	.830	.595	.693	.250	.659	.363	.693	.762	.726	.945	.678	.789	.046	.524	.085	.675	.554
OneCoin	.892	.695	.781	.510	.696	.589	.673	.831	.744	.888	.848	.867	.875	.333	.483	.797	.696
GLAD	.887	.630	.737	.329	.760	.460	.609	.797	.690	.922	.751	.828	.632	.571	.600	.734	.619
M-MSR	.652	.440	.525	.184	.594	.281	.419	.628	.503	.851	.523	.648	.222	.286	.250	.531	.362
MACE	.872	.807	.838	.445	.802	.572	.720	.819	.766	.961	.810	.879	.353	.857	.500	**.811**	.726
Wawa	.823	.511	.631	.235	.696	.352	.516	.757	.614	.897	.628	.739	.889	.381	.533	.633	.491
ZBS	.845	.554	.670	.278	.737	.404	.548	.779	.643	.916	.683	.782	.750	.429	.545	.677	.546
Advanced UI																	
MV	.665	.364	.470	.150	.544	.235	.437	.701	.539	.866	.509	.641	1.000	.048	.091	.517	.349
DawidSkene	.819	.473	.599	.206	.668	.315	.794	.781	.787	.968	.728	.831	.053	.571	.097	.678	.559
OneCoin	.911	.877	.893	.571	.779	.659	.805	.894	.847	.951	.880	.914	.909	.476	.625	**.873**	.811
GLAD	.867	.712	.782	.344	.724	.466	.695	.860	.769	.953	.790	.864	.818	.429	.563	.781	.684
M-MSR	.633	.433	.514	.146	.535	.229	.440	.663	.529	.865	.483	.620	.500	.143	.222	.512	.345
MACE	.889	.910	.899	.538	.816	.648	.835	.871	.852	.967	.858	.909	.415	.810	.548	.869	.807
Wawa	.706	.501	.586	.181	.608	.279	.504	.722	.593	.911	.567	.699	1.000	.238	.385	.586	.437
ZBS	.830	.650	.729	.266	.705	.386	.633	.813	.712	.942	.708	.808	.889	.381	.533	.715	.599
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764

Table 5: All Workers integrated with GPT-4 Table. All models use Bio Expert as the gold standard. Baseline is the Majority Vote (MV).

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI																	
MV	.814	.426	.559	.223	.654	.333	.488	.712	.579	.853	.644	.734	1.000	.286	.444	.609	.451
DawidSkene	.872	.519	.650	.280	.737	.406	.746	.803	.773	.936	.808	.867	.125	.667	.211	.737	.626
OneCoin	.888	.725	.798	.507	.714	.593	.689	.826	.751	.897	.849	.872	.875	.333	.483	.804	.708
GLAD	.876	.669	.759	.354	.760	.483	.644	.806	.716	.928	.775	.845	.609	.667	.636	.757	.649
M-MSR	.756	.536	.627	.225	.627	.331	.511	.697	.590	.886	.647	.748	.900	.429	.581	.630	.482
MACE	.873	.801	.836	.442	.797	.569	.719	.819	.766	.960	.809	.878	.327	.857	.474	**.809**	.724
Wawa	.833	.577	.682	.264	.728	.388	.553	.759	.640	.912	.673	.774	.818	.429	.563	.672	.540
ZBS	.849	.613	.712	.305	.747	.433	.592	.788	.676	.921	.721	.809	.667	.476	.556	.712	.590
Advanced UI																	
MV	.740	.417	.533	.189	.590	.286	.475	.738	.578	.885	.595	.712	1.000	.048	.091	.583	.424
DawidSkene	.874	.663	.754	.311	.756	.441	.833	.860	.847	.972	.804	.880	.126	.762	.216	.782	.690
OneCoin	.908	.891	.899	.579	.797	.671	.814	.884	.848	.952	.881	.915	.909	.476	.625	**.875**	.815
GLAD	.880	.754	.812	.384	.765	.512	.718	.863	.784	.958	.806	.875	.750	.571	.649	.802	.714
M-MSR	.663	.539	.594	.186	.581	.282	.497	.696	.580	.897	.555	.686	.333	.238	.278	.581	.429
MACE	.890	.908	.899	.540	.816	.650	.838	.869	.853	.966	.861	.910	.425	.810	.557	.870	.809
Wawa	.779	.615	.687	.253	.673	.367	.586	.787	.672	.934	.675	.784	1.000	.333	.500	.683	.556
ZBS	.852	.732	.787	.351	.737	.475	.685	.850	.759	.953	.773	.854	.909	.476	.625	.776	.678
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764

Table 6: Exclude-By-Worker integrated with GPT-4 Table. All models use Bio Expert as the gold standard. Baseline is the Majority Vote (MV). From Exclude-By-Worker results, One-Coin Dawid-Skene aggregation model achieves the highest accuracy for both basic and advanced interface. Advanced One-Coin Dawid-Skene reaches 86.6% and outperforms other aggregation models in every aspects. The accuracy from advanced One-Coin Dawid-Skene almost reaches the accuracy of the GPT-4 (t=.2), 82.7%.

Eval	Background			Purpose			Method			Finding			Other			Acc	Kappa
Label	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
Basic UI																	
MV	.852	.732	.787	.351	.737	.475	.685	.850	.759	.953	.773	.854	.909	.476	.625	.776	.678
DawidSkene	.864	.484	.621	.263	.724	.386	.721	.784	.751	.936	.782	.852	.088	.619	.154	.712	.593
OneCoin	.888	.712	.790	.513	.710	.596	.675	.834	.746	.896	.843	.869	.875	.333	.483	.800	.702
GLAD	.873	.629	.731	.328	.765	.459	.608	.797	.690	.924	.744	.825	.684	.619	.650	.731	.615
M-MSR	.689	.491	.574	.202	.581	.299	.479	.656	.553	.858	.609	.713	.500	.333	.400	.590	.428
MACE	.871	.804	.836	.444	.797	.570	.718	.818	.765	.960	.810	.878	.346	.857	.493	**.810**	.724
Wawa	.807	.526	.637	.239	.700	.356	.513	.757	.612	.912	.627	.743	1.000	.381	.552	.636	.495
ZBS	.834	.569	.676	.276	.747	.403	.548	.778	.643	.919	.671	.776	.818	.429	.563	.675	.544
Advances UI																	
MV	.694	.370	.482	.149	.525	.232	.434	.701	.536	.864	.520	.649	1.000	.048	.091	.523	.354
DawidSkene	.838	.556	.668	.242	.700	.360	.792	.784	.788	.969	.765	.855	.077	.667	.139	.718	.607
OneCoin	.909	.884	.896	.581	.793	.671	.808	.893	.848	.953	.879	.914	.909	.476	.625	**.874**	.813
GLAD	.865	.698	.772	.325	.700	.444	.671	.857	.753	.951	.773	.853	.889	.381	.533	.767	.665
M-MSR	.537	.450	.489	.169	.544	.257	.451	.650	.532	.869	.498	.634	.313	.238	.270	.522	.353
MACE	.884	.910	.897	.546	.816	.654	.835	.871	.852	.968	.859	.910	.425	.810	.557	.869	.808
Wawa	.723	.511	.599	.186	.613	.286	.509	.726	.598	.913	.580	.710	1.000	.286	.444	.597	.449
ZBS	.817	.653	.726	.260	.691	.377	.631	.810	.710	.942	.700	.803	1.000	.381	.552	.711	.593
GPT-4 (t=.2)	.860	.913	.885	.499	.843	.627	.775	.871	.820	.982	.784	.872	.322	.905	.475	.836	.764

Table 7: Exclude-By-Batch integrated with GPT-4 Table. All models use Bio Expert as the gold standard. Baseline is the Majority Vote (MV).

Basic Interface

Acc of GPT-4 (t=0.2) =**.836**	All Workers			Exclude-By-Worker			Exclude-By-Batch		
Method	**Acc**	**P-Value**	**95% CI**	**Acc**	**P-Value**	**95% CI**	**Acc**	**P-Value**	**95% CI**
MV	.551	<.001	[.534, .569]	.567	<.001	[.550, .584]	.609	<.001	[.592, .626]
DawidSkene	.675	<.001	[.659, .691]	.712	<.001	[.696, .727]	.737	<.001	[.722, .753]
OneCoin	.797	.002	[.783, .811]	.800	.006	[.786, .814]	.804	.002	[.790, .818]
GLAD	.734	<.001	[.719, .749]	.731	<.001	[.715, .746]	.756	<.001	[.741, .771]
M-MSR	.531	<.001	[.513, .548]	.590	<.001	[.572, .607]	.630	<.001	[.614, .647]
MACE	**.811**	.001	**[.797, .824]**	**.809**	<.001	**[.796, .823]**	**.809**	.001	**[.795, .823]**
Wawa	.633	<.001	[.617, .650]	.636	<.001	[.619, .653]	.672	<.001	[.656, .689]
ZBS	.677	<.001	[.661, .694]	.675	<.001	[.659, .691]	.712	<.001	[.696, .727]
Avg. Acc	.676	-	-	.716	-	-	.690	-	-
#workers	216			134			176		

Table 8: Aggregation Accuracy Results of the Basic Interface integrated with GPT4 Group. Bold and underline highlight the highest score within the column and across the table, respectively. P-value is obtained by comparing with GPT-4 over the article-level accuracy. (**: p<0.01; ***: p<0.001. Paired t-test. N=200)

Advanced Interface

Acc of GPT-4 (t=0.2) =**.836**	All Workers			Exclude-By-Worker			Exclude-By-Batch		
Method	**Acc**	**P-Value**	**95% CI**	**Acc**	**P-Value**	**95% CI**	**Acc**	**P-Value**	**95% CI**
MV	.517	<.001	[.500, .535]	.583	<.001	[.565, .600]	.523	<.001	[.505, .540]
DawidSkene	.678	<.001	[.662, .695]	.782	<.001	[.767, .796]	.718	<.001	[.702, .734]
OneCoin	**.873**	.002	[.861, .884]	**.875**	.001	[.864, .887]	**.874**	<.001	[.863, .886]
GLAD	.781	<.001	[.763, .792]	.802	<.001	[.788, .816]	.767	<.001	[.753, .782]
M-MSR	.512	<.001	[.494, .529]	.581	<.001	[.564, .599]	.522	<.001	[.504, .539]
MACE	.869	.004	[.857, .880]	.870	.010	[.858, .881]	.869	.003	[.858, .881]
Wawa	.586	<.001	[.569, .604]	.683	<.001	[.667, .700]	.597	<.001	[.580, .614]
ZBS	.715	<.001	[.700, .731]	.776	<.001	[.762, .791]	.711	<.001	[.695, .727]

Table 9: Aggregation Accuracy Results of the Advanced Interface integrated with GPT4 Group. Bold and underline highlight the highest score within the column and across the table, respectively. OneCoin and MACE are only two aggregation methods that outperform GPT-4 and the differences are statistically significant, shown in the table. P-value is obtained by comparing with GPT-4 over the article-level accuracy. (**: p<0.01; ***: p<0.001. Paired t-test. N=200)

Zero-shot Prompt
Classify the given text into one of the following labels.

[Background]: Text segments answer one or more of these questions: Why is this problem important?, What relevant works have been created before?, What is still missing in the previous works?, What are the high-level research questions?, How might this help other research or researchers?
[Purpose]: Text segments answer one or more of these questions: What specific things do the researchers want to do?, What specific knowledge do the researchers want to gain?, What specific hypothesis do the researchers want to test?
[Method]: Text segments answer one or more of these questions: How did the researchers do the work or find what they sought?, What are the procedures and steps of the research?
[Finding]: Text segments answer one or more of these questions: What did the researchers find out?, Did the proposed methods work?, Did the thing behave as the researchers expected?
[Other]: Text fragments that do NOT fit into any of the four categories above. Text fragments that are NOT part of the article. Text fragments that are NOT in English. Text fragments that contains ONLY reference marks (e.g., "[1,2,3,4,5") or ONLY dates (e.g., "April 20, 2008"). Captions for figures and tables (e.g. "Figure 1: Experimental Result of ...", or "Table 1: The Typical Symptoms of ...") Formatting errors. I really don't know or I'm not sure.

FAQs
1. This text fragment has terms that I don't understand. What should I do? Please use the context in the article to figure out the focus. You can look up terms you don't know if you feel like you need to understand them.
2. This text fragment is too short to mean anything. What should I do? If the text fragment is too short to have significant meanings, you could consider the entire sentence and answer based on the entire sentence.
3. This text fragment is NOT in English. What should I do? If the whole fragment (or the majority of words in the fragment) is in Non-English, please label it as "Other". If the majority of the words in this fragment are in English with a few non-English words, please judge the label normally.
4. I'm not sure if this should be a "background" or a "finding." How do I tell? When a sentence occurs in the earlier part of an article, and it is presented as a known fact or looks authoritative, it is often a "background" information.
5. Do "potential applications of the proposed work" count as "background" or "purpose"? It should be "background." The "purpose" refers to specific things the paper wants to achieve.
6. If the article says it's a "literature review" (e.g., "We reviewed the literature" / "In this article, we review.." etc), would we classify those as finding/contribution or purpose? Most parts of a literature review paper should still be "background" or "purpose", and only the "insight" drew from a set of prior works can be viewed as a "finding/contribution".
7. What should I do with the case study on a patient? Typically, it has a patient come in with a set of signs and symptoms in the ER, and then the patient gets assessed and diagnosed. The patient is admitted to the hospital ICU and tests are done and they may be diagnosed with something else. In such cases, please label the interventions done by the medical staff (e.g., CT scans, X-rays, and medications given) as "Method", and the patient's final result (e.g. the patient's pneumonia resolved and he was released from the hospital) as "Finding/Contribution".

Classify the following sentence into one of the label: Background, Purpose, Method, Finding, and Other.
Provide answer in format of "'fragment-i []"'
fragment-1 Text: "'{**Sentence-1**}"'
Label: []
fragment-2 Text: "'{**Sentence-2**}"'
Label: []
fragment-3 Text: "'{**Sentence-3**}"'
Label: []
.....

Table 10: Zero-shot prompt used when calling GPT-4. The {Sentence-n} will be replaced by the following sentence in the abstract we would like to predict.