Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

On the Role of Large Language Models in Crowdsourcing Misinformation Assessment

Jiechen Xu, Lei Han, Shazia Sadiq, Gianluca Demartini

The University of Queensland, Australia
jiechen.xu@uq.net.au, l.han@uq.edu.au, shazia@eecs.uq.edu.au, g.demartini@uq.edu.au

Abstract

The proliferation of online misinformation significantly undermines the credibility of web content. Recently, crowd workers have been successfully employed to assess misinformation to address the limited scalability of professional fact-checkers. An alternative approach to crowdsourcing is the use of large language models (LLMs). These models are however also not perfect. In this paper, we investigate the scenario of crowd workers working *in collaboration with* LLMs to assess misinformation. We perform a study where we ask crowd workers to judge the truthfulness of statements under different conditions: with and without LLMs labels and explanations. Our results show that crowd workers tend to overestimate truthfulness when exposed to LLM-generated information. Crowd workers are misled by wrong LLM labels, but, on the other hand, their self-reported confidence is lower when they make mistakes due to relying on the LLM. We also observe diverse behaviors among crowd workers when the LLM is presented, indicating that leveraging LLMs can be considered a distinct working strategy.

Introduction

The spread of misinformation online has been creating increasing concern for society. The risk is that organized efforts to spread misinformation could mislead large parts of the population and drive social behavior in desired directions. There have been different approaches proposed to deal with this growing challenge. For example, researchers have been developing effective methods to automatically detect online misinformation (Gatto, Basak, and Preum 2023; Weinzierl, Hopfer, and Harabagiu 2021; Shu, Wang, and Liu 2019). This is used to support expert fact-checkers who would be otherwise overwhelmed by the amount of content to be fact-checked. An interesting alternative to using experts or ML to deal with misinformation is the use of crowdsourcing. Recent research has looked at the feasibility of crowdsourcing as a way to detect and label misinformation (La Barbera et al. 2020; Roitero et al. 2020).

More recently, the introduction of Large Language Models (LLMs) has triggered a new line of research looking at both (i) the way LLMs can do tasks in place of humans as well as (ii) how human behavior is affected by the interaction with LLMs. Examples include mathematical formalization (Wu et al. 2022), news summarization (Zhang et al. 2023), or coding (Xu et al. 2022a). While powerful, existing limitations of LLMs include the imperfect output generated by LLMs also including completely made-up information resulting from the generative process.

In this paper, we look at the impact of LLMs supporting crowd workers assessing the truthfulness of political statements. We are specifically interested in the dimensions of *trust* and *reliance* in the LLMs used by crowd workers as co-pilots in their judgment tasks. We analyze if and how some crowd workers believe that the LLM provides accurate judgments (*trust*) more than others and if and how some crowd workers make use of LLM recommendations (*reliance*) more than others.

Our research questions include the following:

- **RQ1:** Does presenting LLM-generated outputs (truthfulness labels and explanations) have an effect on the quality of assessments from crowd workers?
- **RQ2:** What is the effect on crowd workers' confidence level to their assessments from LLM-generated outputs?
- **RQ3:** What is the impact of LLM-generated outputs on crowd workers in terms of the reliance on LLM and the subjective trust in LLM.
- **RQ4:** How does LLM-generated output affect crowd workers' behaviors when they assess misinformation?

We observed a strong influence of GPT-3.5 on crowd worker decisions. Our findings show that crowd workers tend to overestimate truthfulness when exposed to LLM-generated information.

Related Work

Crowdsourcing for Misinformation Detection

The spread of online misinformation has raised concerns about how the public can be informed and make meaningful decisions in a democratic society (Kumar, West, and Leskovec 2016). To deal with this issue expert human fact-checkers with a background in journalism follow a forensic process to debunk misinformation by providing evidence and explanations to news readers. However, given the vast amount of misinformation to be fact-checked, human experts face growing challenges. To this end, automated meth-

ods have been developed (Guo, Schlichtkrull, and Vlachos 2022).

To deal with the challenges of fact-checking and with the limitations of automated methods in terms of accuracy and possible bias, crowdsourcing has been considered as an alternative to automated methods given its ability to scale to large amounts of data. Roitero et al. (2020) looked at how crowd workers compare with expert fact-checkers highlighting a good level of agreement between the two. Later, Roitero et al. (2021) looked at the longitudinal dimension of crowdsourced truthfulness assessment observing a consistency in the generated labels. La Barbera et al. (2020) observed a political bias in crowd-generated truthfulness labels. In summary, the key conclusion that can be drawn from this body of work is that the crowd can be a viable alternative to experts or to automatically generate misinformation labels with the risk of potential bias.

As compared to the existing literature, in this paper, we focus on understanding the impact of text generated by LLMs in support of crowd workers assessing the truthfulness of political claims. Related to this, there is a study that has looked at the impact of presenting crowd workers assessing truthfulness with external information about how other workers have judged the same claim (Xu et al. 2022b).

External Input in Micro-task Crowdsourcing

To have a better understanding of crowd workers and potentially boost the outcome of crowdsourcing tasks, a number of studies have been conducted to look at expanding the current micro-task design by adding external information to crowd workers. Zhu et al. (2014) embed reviews from peer crowd workers to varied types of crowdsourcing tasks and aim to optimize the quality of task outcomes. The study implies that the strategy of including reviews should be contingent on the level of subjectivity inherent in the task. Similarly, Lekschas et al. (2021) employ crowd workers to comment on the works created by artists who are able to continuously improve their quality of artworks after viewing the crowd-based comments. Doroudi et al. (2016) conduct a study towards enhancing the quality of a problem-solving task by showing novice crowd workers with example solutions crafted by either experts or peer crowd workers. The result indicates new crowd workers can develop effective task-completion strategies from given examples. Hettiachchi et al. (2021) underscore how utilizing 'visible gold' (i.e., questions that offer feedback to workers as indicators of accuracy) can significantly improve the quality of work in a crowdsourced face detection task. In the context of relevance assessment, Eickhoff (2018) enrich the task design by implanting assessment results from previous crowd workers. The results indicate a significant bandwagon effect, that is, crowd workers tend to exclusively rely on the provided answers to complete the task. Similarly, Xu et al. (2023) investigate the effect of human and machine metadata in the crowdsourced relevance assessment. They have identified a strong preference for crowd workers on human metadata and showed the connection between the metadata quality and the assessment quality. Duan, Ho, and Yin (2022) apply a similar approach that adding information either generated by humans or machines to a recidivism risk evaluation task to explore the impact of changing task design with respect to crowd workers' performance. Heuer and Glassman (2022) provide online fact-checkers with a checklist that aims to assist fact-checkers. They find that the usefulness of the checklist is non-trivial. This body of literature has comprehensively investigated the impact of such external information on crowd workers. Our work aims to expand existing research by involving LLMs as the source of information.

Human-LLM Interaction

LLMs have recently received a lot of attention from the research community as well as from the general public. With the term LLM, we typically refer to transformer models pretrained with large amounts of documents from which tokens are extracted in which the decoder is used to generate natural language text. Popular examples of LLMs include generative pre-trained transformer (GPT) models (Brown et al. 2020). While significant work has been conducted to understand the effectiveness of these models on a large number of human tasks (Zhang et al. 2023), in this work, we instead focus on how these LLMs may impact human decision-making processes when human decision-makers are exposed to LLM output. Related to the topic of human-AI interaction, recent work as looked at how human decision-makers rely on AI. For example, Jahanbakhsh et al. (2023) explore the use of a personalized AI that learns from individual user assessments and predicts how the same user would assess other content for identifying misinformation on social media. A user study shows that the AI's predictions influence users' judgment, but this influence can be reduced when users provide reasoning for their assessment. Epstein et al. (2022) look at the effectiveness of AI-generated warnings on social media platforms to combat misinformation. They find that providing explanations for the warnings can increase users' discernment, but the effectiveness can be undermined by various factors (e.g., user attention and education level). Vasconcelos et al. (2023) investigate the effect of AI-generated explanations to reduce users' over-reliance on AI and show the scenarios that explanations can mitigate over-reliance.

Methodology

Dataset

In this study, we leverage a dataset containing over 12,000 political statements from a well-known political fact-checking website: *PolitiFact* (Wang 2017). Statements in this dataset are mainly made by US politicians belonging to either the Democrat Party or the Republican Party, evaluated for truthfulness by PolitiFact's editors using a six-level scale: *(0) pants-on-fire, (1) false, (2) barely-true, (3) half-true, (4) mostly-true* and *(5) true*. From this dataset, we then select a subset of 120 statements that are balanced in terms of the two main parties ($60*2$) and truthfulness level ($20*6$). This subset of statements has also been used by related work that looked at crowdsourcing misinformation assessments (Roitero et al. 2020; Draws et al. 2022; Xu et al. 2022b).

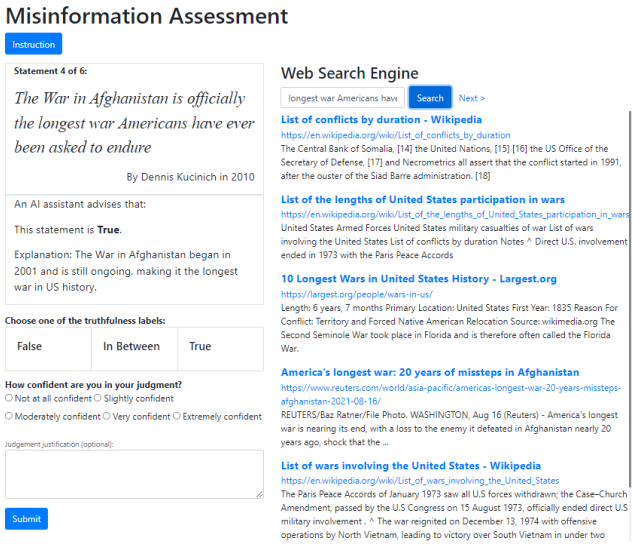


Figure 1: Task interface for the condition Label+Exp.

To investigate the impact of LLMs on crowdsourcing misinformation assessment, we utilize the GPT-3.5 model developed by OpenAI (Brown et al. 2020). Specifically, we employ the *'text-davinci-003'* variant of the model to evaluate the truthfulness of statements from the *PolitiFact* dataset. Via prompting the model, we explicitly request two types of output from GPT-3.5 for each statement: (i) a truthfulness label (from 0 to 5, same as *PolifiFact*) and (ii) a short natural language explanation to justify the decision in a fixed format. Note that the explanations may not directly reveal the corresponding truthfulness label produced by the model. We set parameters 'temperature' and 'top-p' to 0 and 1.0 to control GPT-3.5 to generate consistent responses. Here is an example of GPT-3.5-generated results:

> **Statement**: The War in Afghanistan is officially the longest war Americans have ever been asked to endure.
> **Speaker**: Dennis Kucinich
> **Year**: 2010
> **Truthfulness label**: 5
> **Explanation**: The War in Afghanistan began in 2001 and is still ongoing, making it the longest war in US history.

By merging the LLM-generated labels and editorial labels into binary (mapping *PolitiFact* labels 0-2: true, 3-5: false), we observed a model accuracy rate of approximately 0.68. This level of effectiveness is comparable to a recent study conducted by Hoes, Altay, and Bermeo (2023).

Experimental Conditions

We run an experiment using a 2×2 factorial design based on the presence of two kinds of LLM-generated data (i.e., truthfulness label and explanation). Hence, our experiment contains 4 conditions:

- *Baseline*. In this condition, we do not provide any kind of LLM-generated answer, which serves as the baseline in our study and is comparable to the design used in the literature on crowdsourced misinformation assessments.
- *Label*. In this condition, crowd assessors are exposed to LLM-generated truthfulness labels, presented on the same scale as the labels they are required to provide.
- *Explanation*. This condition contains solely the natural language explanations generated by GPT-3.5.
- *Label+Exp*. For this condition, we show both labels and explanations to crowd workers.

Crowdsourcing Task Design

We divide the selected 120 political statements into 20 tasks/units each of them containing 6 statements. To ensure all units are balanced in terms of truthfulness level and political parties to mitigate the cognitive bias of crowd workers (La Barbera et al. 2020; Draws et al. 2022), we merge the original 6-level truthfulness scale into a 3-level one (*true, in between and false*) based on the observations made by Roitero et al. (2020). The set of labels given by PolitiFact editors is considered the ground truth in this study. In addition, we compose each unit using 4 statements that are correctly labeled by GPT-3.5, while 2 statements were incorrectly labeled to mimic the actual GPT-3.5's accuracy on our dataset. The order of the 6 statements is then shuffled for every crowd worker and every task to remove potential learning effects (He, Kuiper, and Gadiraju 2023).

Crowd assessors are recruited from the online crowdsourcing platform *Prolific*, which has been shown as an effective choice for running complex human assessment studies (Xu, Zhou, and Gadiraju 2020). To conduct this study, we recruit participants who are physically located in the USA and use English as their first language. Additionally, potential participants are required to meet certain criteria, including an approval rate of at least 80% and a minimum completion of 10 prior tasks on the platform. These criteria are regarded as a method to obtain reliable crowd assessors. Crowd workers will be assigned randomly to one task unit in one of the four experimental conditions, and participating multiple times is not allowed. Prior to starting the task, participants were required to review an information sheet explaining the data collection process and the intended use of the collected data and consent to it[1]. Then, participants will be redirected to an external page hosted by the authors' institution to continue the task. Figure 1 shows the task interface for one of the experimental conditions, which comprises the following elements on the left-hand side: (a) an instruction button, (b) the statement information including speaker and year, (c) GPT-3.5's output (varies across conditions), (d) radio buttons for the workers to indicate their confidence levels and (e) a text box for an optional justification text input. On the right-hand side, we provide a customized search engine[2] that allows crowd assessors to find evidence to support their assessments (Roitero et al. 2020). A search result consists of the title of the webpage, the URL, and a snippet of text

[1]This study has been approved by the authors' institution IRB.

[2]The search engine leverages Bing APIs and does not return any result from the PolitiFact website as crowd assessors may otherwise be able to directly retrieve the ground truth.

from the web page which may contain the query terms. Note that the initial state of the task interface does not include any search result or any search query placeholder.

In addition to assessing the 6 political statements, participants are requested to complete a pre-task questionnaire (ATI, TIA-PtT) and an exit questionnaire (TIA-Trust). To prevent crowd workers from submitting random answers, we embed attention-check questions throughout the task session. Submissions from workers who failed to answer any attention-check question are not used in data analysis. We compensate participants €1.2 per task based on the average task completion time measured during a pilot study and the Prolific recommended hourly rate.

Measures

Measuring Assessment Quality. To address RQ1, we leverage various metrics to evaluate the quality of labels given by crowd workers compared to the labels from PolitiFact's editors (ground truth). First, we consider two kinds of *errors* between crowd workers and the editors: E_{edit} and AE_{edit}. E_{edit} represents the extent of disparity (ranges in $[-2, 2]$) between labels provided by a crowd worker and an editor for a given statement, indicating the presence of overestimation (positive values) or underestimation (negative values) tendencies. AE_{edit} quantifies the magnitude of annotation bias by calculating the absolute error between labels assigned by crowd workers and editors, with values ranging from 0 to 2. This set of error-based evaluation was also used by Draws et al. (2022). We also consider to evaluate the inter-rater agreement of collected labels (Checco et al. 2017). Here, we apply two types of agreement: (i) *external agreement*, which measures the agreement between labels from crowd workers and the ground truth, and (ii) *internal agreement*, which assesses the agreement among crowd workers working on the same set of statements. Specifically, we utilize Krippendorff's α (Krippendorff 2011) to calculate both types of agreement which has value range from -1 (completely disagree) to 0 (random) to 1 (completely agree).

Measuring Reliance and Trust. Reliance is measured by analyzing the alignment of labels generated by the LLM and crowd workers. With a similar idea to measuring quality, we embed *error-based metrics* including E_{LLM} and AE_{LLM} which compute the error and absolute error between crowd workers and LLM, respectively. In addition, we apply *Agreement Fraction* (He, Kuiper, and Gadiraju 2023), which calculates the rate of crowd workers' decisions that agree with the LLM's advice. Following previous research (Tolmeijer et al. 2022), we consider two validated questionnaires using Likert scales to measure the subjective trust in LLM: *Trust in Automation* (TiA) (Körber 2019) and *Affinity for Technology Interaction Scale* (ATI) (Franke, Attig, and Wessel 2019). For TIA, we apply two sub-scales related to our study: *Propensity to Trust* (TiA-PtT), *Trust in Automation* (TiA-Trust).

Confidence and Behavioral Indicators. In our study, we apply *self-reported confidence* as the metric of confidence of crowd workers in their assessments, using a 5-level Likert scale following prior research (Qu et al. 2022) to address RQ3. To study behavioral aspects (RQ4), we consider the time usage and the interaction with the search engine as behavioral indicators in this work. Given the nature of the misinformation assessing task, making use of a search engine can be regarded as the proxy to measure workers' efforts (Roitero et al. 2020). We particularly examine the *Number of Queries* as the evidence of the extent to which workers actively engage with the search engine.

	Completed	**Abandoned**	**Failed**
Baseline	60 (58.8%)	30 (29.4%)	12 (11.8%)
Label	60 (53.6%)	45 (40.2%)	7 (6.2%)
Explanation	60 (51.3%)	44 (37.6%)	13 (11.1%)
Label+Exp	60 (56.6%)	33 (31.1%)	13 (12.3%)

Table 1: Crowd worker participation rates.

Results

Demographics

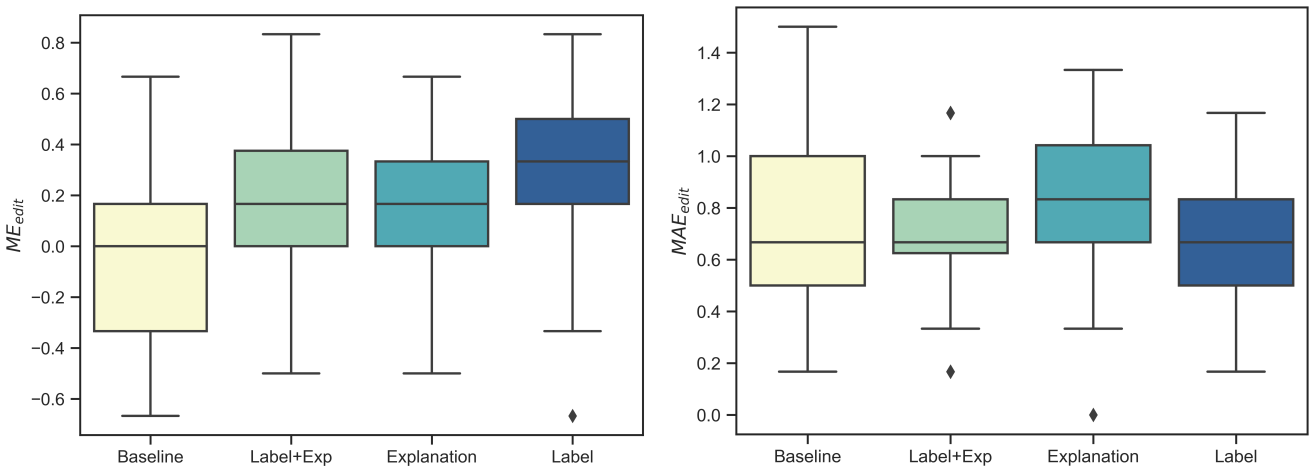
In this study, we recruited 437 crowd workers through the Prolific platform. Table 1 illustrates the number of workers who completed the task, abandoned the task, and failed the attention check questions across experimental conditions. Out of all the participants, 46% identified as female, while 51% identified as male. The remaining participants chose not to disclose their gender identity. The average age of the participants is found to be 36 years, with a standard deviation of 13, indicating a diverse range of age groups. We found that most workers were well-educated since 63% of them claim that they acquired a 4-year college degree or above. To ensure that demographic variables (e.g., age and education level) have no effect to the dependent variables (e.g., ME_{edit}) in our study, we leverage *Spearman correlation tests* and *Pearson correlation tests* to relate demographic variables to the dependent variables included in results section. All tests resulted in $p > 0.05$, which indicates we can ignore demographic variables in our analysis.

RQ1: Quality of Assessments

Next, we report about the effects of LLM-generated answers on crowd worker assessment quality. To test, we propose the following Null Hypothesis for this section:

> **Null Hypothesis 1** *LLM-generated answers (i.e., labels and explanations) do not have an effect on assessment quality.*

Overestimated Truthfulness Impacted by LLM. First, we consider the crowd workers' judgment mean error as compared to experts ME_{edit}, which looks at the mean difference between labels from crowd workers labels and PolitiFact's editors. Figure 2a shows the distribution of ME_{edit} over four experimental conditions. We notice that the median value for condition Baseline is around 0 while other three conditions have higher median values. To examine



(a) (b)

Figure 2: ME_{edit} and MAE_{edit} over conditions.

Condition	% Over	% Under	Accuracy
Baseline	27.50	30.83	41.67
Label	38.06	18.61	43.33
Explanation	36.11	25.28	38.61
Label+Exp	37.78	20.56	41.67

Table 2: Percentages of assessments for overestimation, underestimation and accuracy across conditions.

the effect of showing labels and/or explanations generated by GPT-3.5, an *aligned ranks transformation ANOVA* (ART ANOVA) (Wobbrock et al. 2011) is applied due to the data in Fig. 2a being not normally distributed. The ART ANOVA is a non-parametric test designed to analyze multiple independent variables and their interactions, offering flexibility in capturing complex relationships and measures within the data. The result of ART ANOVA shows a significant interaction effect between the two factors ($F = 10.8796, p = 0.0011$, partial $\eta^2 = 0.0440$). A post-hoc pair-wise comparison using Tukey's HSD adjustment shows condition Baseline has significantly lower ME_{edit} compared to conditions having LLM-generated information ($p < .01$ for Label, Explanation and Label+Exp). Within conditions Label, Explanation and Label+Explanation, we do not find any of them having ME_{edit} significantly different to others, while we observe that Label condition has a higher median as shown in Fig. 2a. Table 2 shows the distribution of assessments being overestimated, underestimated, and accurate across different conditions. It is evident that crowd workers tend to overestimate truthfulness when exposed to LLM-generated information and hence, Null Hypothesis 1 should be rejected. Conversely, workers who do not have an access to GPT-3.5's assistance demonstrate a higher rate of underestimation errors. In term of accuracy, the values of Baseline, Label and Label+Exp are comparable, while condition Explanation achieves the lowest value due to an increase in underestimation errors.

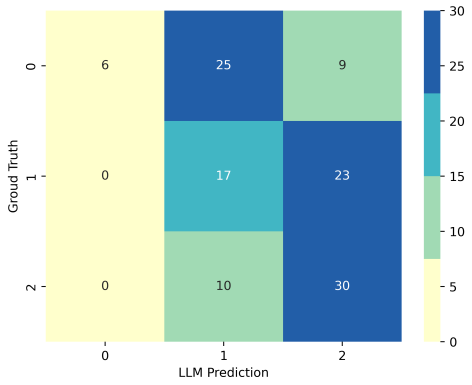


Figure 3: Confusion matrix for numbers of assessments by GPT-3.5 against the ground truth labels. Labels for row and column are ground truth labels and GPT-3.5's labels, respectively. Notation: 0 – false, 1 – in-between, 2 – true.

To have a better understanding on the reason why crowd workers tend to overestimate the truthfulness, we report the judging performance of GPT-3.5 on the same statement corpus. We calculate the mean error between GPT-3.5 and the ground truth (E_{LLM_edit}), which results in $M \pm SD(E_{LLM\_edit}) = 0.5 \pm 0.22$. This observation indicates that GPT-3.5 tends to overestimate truthfulness labels, which is aligned with the recent results in the literature showing how LLMs are more effective when assessing true statements than false ones (Hoes, Altay, and Bermeo 2023). Figure 3 shows presents the number of labels given by GPT-3.5 for different truthfulness levels in the ground truth. From the top left to the bottom right, the three cells on the diagonal represents the labels that GPT-3.5 is aligned with ground truth. Simultaneously, the upper section of the matrix dis-

Conditions	External M	External SD	Internal M	Internal SD
Baseline	0.24	0.38	-0.05	0.39
Label	0.27	0.35	-0.00	0.40
Explanation	0.14	0.38	-0.04	0.37
Label+Exp	0.27	0.31	-0.04	0.39

Table 3: Mean and standard deviation of Krippendorff's α over conditions.

Variables	External F	External p	Internal F	Internal p
Label	3.22	0.07	0.20	0.65
Explanation	1.13	0.29	0.06	0.80
Label×Exp	1.01	0.31	0.29	0.59

Table 4: ANOVA results of Krippendorff's α for external agreement and internal agreement over conditions.

plays the count of judgments made by GPT-3.5 that have overestimated truthfulness, while the lower section contains judgments with underestimations of truthfulness. It is evident that GPT-3.5 tends to exhibit a bias towards overestimating the truthfulness of this set of political statements, which subsequently appears to have influenced the crowd workers' assessments. These findings explains why crowd workers who are exposed to LLMs data make more overestimation errors.

Figure 2b shows the the distributions of MAE_{edit} across conditions. By running ART ANOVA, neither interaction effect nor main effect for each factor is identified ($p = 0.6172$ for label×explanation, $p = 0.062$ and $p = 0.078$ for label and explanation, respectively). Crowd workers who are exposed to LLM-generated answers make more overestimation errors. However, it is counter intuitive that no similar observation can be found via analyzing MAE_{edit}. This indicates how the level of annotation bias among all experimental conditions are comparable. A possible explanation is that crowd workers in Baseline are making more underestimation errors (see Tab. 2), which generates a similar bias level compared to other conditions.

Internal and External Agreement. We now turn to analyze crowd workers' assessment quality by computing internal and external agreement. We use Krippendorff's α to measure agreement. The results are summarized in Table 3. In terms of external agreement, we can observe that Label and Label+Exp lead to best judgment quality with respect to mean α and, condition Explanation leads to lowest quality. After conducting tests for normality and homogeneity of variances, we apply *two-way ANOVA* on external agreement and internal agreement. The results for ANOVA are shown in Table 4. In summary, we have not found any significant effect from LLM-generated labels and explanations in terms of external agreement among these four conditions. As an important measure for the quality of labels acquired from crowd workers (Checco et al. 2017), our findings with regard to external agreement show that being exposed to data from a LLM does not have an impact on crowd workers' judgment qualities . On the other hand, the analysis in the prior section indicates a gap between the quality of labels collected in the Baseline condition and the LLM-supported conditions. This contradiction can also be the result of more underestimation errors appearing in the Baseline condition.

On the other hand, we notice that the mean Krippendorff's α for internal agreement is fairly low across all experimental conditions. This result is consistent with previous research conducted on crowdsourcing misinformation assessment (Roitero et al. 2020; Xu, Zhou, and Gadiraju 2020), even though we use a different crowdsourcing platform. We also did not identify any condition that is statistically significantly different to others based on internal agreement. This indicates that providing answers from a LLM does not improve the consistency of labels generated by crowd workers.

RQ2: Self-reported Confidence

Next, we explore the impact of LLM-generated answers on crowd workers' self-reported confidence in their judgment. To this end, we test the following Null Hypothesis:

> **Null Hypothesis 2** *LLM-generated answers do not have an effect on self-reported confidence levels.*

For each statement, we ask participants to choose their level of confidence on a Likert scale from 0 (completely unconfident) to 4 (completely confident). To this end, we compute the average confidence level for each crowd worker. Again, we utilize ART ANOVA to examine potential significant differences among the four experimental conditions. The result shows that neither the interaction effect ($F = 0.0259, p = 0.8722$, partial $\eta^2 = 0.0011$) nor the main effect for two independent variables (label: $F = 0.0060, p = 0.9381$, partial $\eta^2 = 0.0002$; explanation: $F = 3.4619, p = 0.0640$, partial $\eta^2 = 0.0145$) are significant, inferring that neither kind of LLM's answer has an effect on of crowd workers' self-reported confidence levels. Given this result, Null Hypothesis 2 can be accepted. In the following, we consider the relation between assessment quality and self-reported confidence level. Here, we utilize external agreement as the metric for evaluating assessment quality. Figure 4 shows the linear regressions conducted for each of the four experimental conditions. We observe a positive correlation between external agreement and mean confidence ($\beta = 0.8425$, $t(58) = 2.708$, $p = 0.009$) in condition Label+Exp. A follow up analysis of variance (ANOVA) shows that external agreement has a statistically significant effect on mean confidence ($F(1, 58) = 7.331$, $p < 0.01$). The regression suggests that under Label+Exp condition, crowd workers exhibit higher levels of confidence when they have answers that are more aligned with the ground truth. This shows how data generated by the LLM may reinforce pre-existing believes in the human subjects participating to our study.

RQ3: Reliance and Trust

Next, we investigate the extent to which crowd workers rely and trust GPT-3.5's answers in different conditions. To this end, we propose two Null Hypotheses to test:

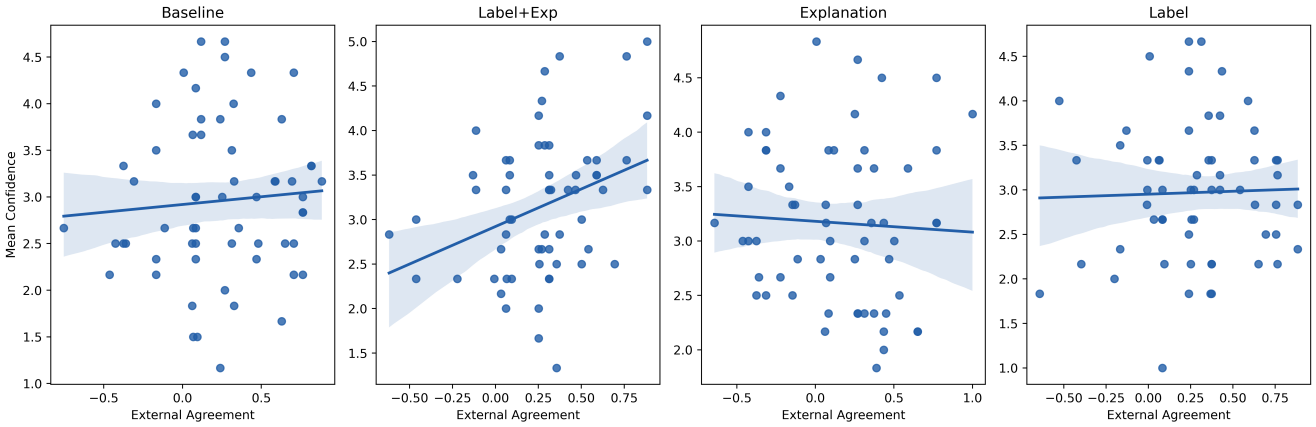


Figure 4: Linear regressions of external agreement and mean confidence among experiment conditions. From left to right: *Baseline*, *Label+Exp*, *Explanation* and *Label*.

Null Hypothesis 3 *The presence of LLM-generated answers does not affect the level of* ***reliance*** *of crowd workers on the LLM.*

Null Hypothesis 4 *The presence of LLM-generated answers does not impact the* ***trust*** *that crowd workers have in the LLM.*

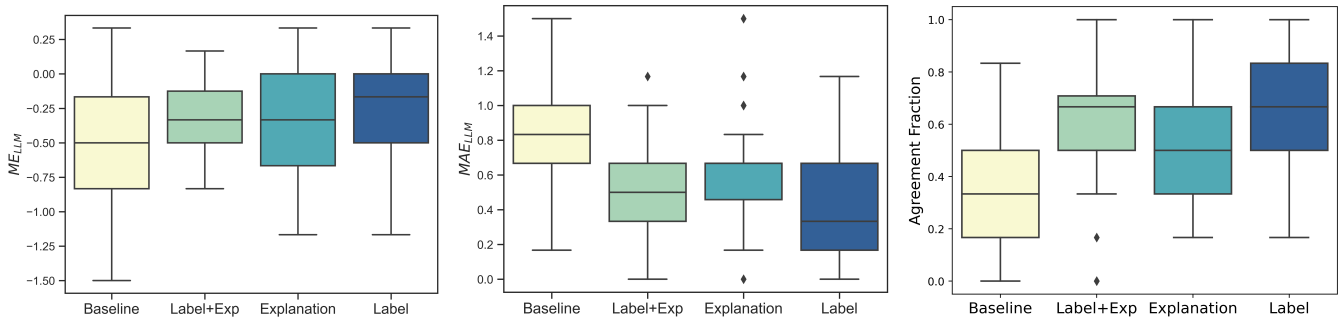
Reliance on LLM's Advice. We first look at the relationship between labels provided by crowd workers and GPT-3.5 for which we start with computing ME_{LLM} (i.e., mean error between crowd workers and GPT-3.5), and MAE_{LLM} (i.e., mean absolute error between crowd workers and GPT-3.5). Then, we also measure the reliance of crowd workers on the LLM using *Agreement Fraction*.

Figure 5a shows crowd workers' ME_{LLM} across four experimental conditions. Note that condition Baseline is presented in Fig. 5a for comparison purposes only as its interface does not include any information generated by GPT-3.5. This situation also applies to other measures reported in this section. By looking at the value distributions in Fig 5a, we can observe how the majority of crowd workers exhibit a tendency to assign lower truthfulness labels relative to the GPT-3.5's assessments (e.g., assigning labels of 0 (i.e., false) or 1 (i.e., in between) while GPT-3.5 assigning a label of 2 (i.e., true)). We conduct a ART ANOVA to investigate the effects on ME_{LLM} and, the result reveals a significant interaction effect ($F = 7.5526, p = 0.0065$, partial $\eta^2 = 0.3010$). The post-hoc pair-wise test (Tukey's HSD) shows that condition Baseline significantly differs form the rest of conditions in terms of ME_{LLM} ($p < 0.01$ for Label and Label+Exp, $p < 0.05$ for Explanation).

Figure 5b shows the distribution of MAE_{LLM} over conditions. Without considering the direction of differences between labels from crowd workers and GPT-3.5, crowd workers achieve smaller mean absolute errors in conditions presenting GPT-3.5's responses. The same set of statistical tests identify a significant interaction effect on MAE_{LLM} ($F = 10.3143, p = 0.0015$, partial $\eta^2 = 0.0419$) and then, condition Baseline is significantly different from other conditions ($p < 0.01$ for all comparisons). We also have discovered significant differences between condition Explanation and the other two conditions containing LLM-generated data ($p < 0.01$ for Label, $p < 0.05$ for Label+Exp). Figure 5c shows the distributions of Agreement Fraction scores across conditions. Form this we can draw similar observations as from Fig. 5b in an opposite direction: Crowd workers obtain higher Agreement Fraction in conditions in which crowd workers have access to LLM's advice. Condition Explanation reaches significantly lower values of Agreement Fraction as compared to Label and Label+Exp (verified by ART ANOVA and post-hoc Tukey's HSD test).

These observations indicate a strong influence of GPT-3.5 on crowd worker decisions, which rejects Null Hypothesis 3. When either LLM-generated labels or explanations are available, labels generated by crowd workers are more aligned with those of the LLM indicating a strong level of reliance. In other words, crowd worker judging behaviour exhibits a bandwagon effect when GPT-3.5 comes in to assist. We also notice that the bandwagon effect is present in condition Explanation although it is not as evident as in condition Label or Label+Exp, indicating that this effect may be mitigated by not explicitly revealing the truthfulness labels generated by the model.

Trust in the LLM. During the task, participants are asked to complete questionnaires which are used as the measurement of subjective trust in the LLM. For the post-questionnaire (TiA-Trust), the questions are rephrased (e.g., from '*I trust the system.*' to '*I believe the AI is judging the truthfulness of statements correctly.*') to highlight to crowd workers that they are reporting about their trust in the LLM. Note that for condition Baseline, we do not ask workers to answer the trust questionnaires. As trust data is not available for the Baseline condition and one of the remaining distributions lacks normality, we apply the *Kruskal-Wallis H Test* to examine the effect of showing LLM-generated answers on trust scores. The obtained result reveals that there is no significant difference among the three conditions



(a) (b) (c)

Figure 5: (a) ME_{LLM}, (b) MAE_{LLM} and (c) Agreement Fraction over experiment conditions.

($H = 0.3190, p = 0.8526$), implying that the variations in LLM-generated answers do not significantly influence the level of trust of crowd workers towards the AI system. This indicates that Null Hypothesis 4 is acceptable. With the observation that self-reported confidence has a positive correlation with external agreement for condition Label+Exp, we are motivated to explore whether trust scores have a correlation with confidence scores. Based on the analysis of regression models we can confirm a positive correlation between self-reported confidence scores and trust scores (ANOVA, $F(1, 58) = 5.4020, p = 0.0237$) only in condition Explanation. By exclusively presenting crowd workers with natural language explanations generated by the LLM, we observe an increase in their confidence levels when they report trusting the generative AI system. However, we do not observe any other correlation between trust scores and other measurements (e.g., external agreement).

Prior work by Tolmeijer et al. (2021) looking at trust formation in the context of AI used as an assistant to complete certain tasks (e.g., searching for suitable apartments) claims that the first impression is vital in terms of building trust in the system. Keeping this in mind, we investigate whether it is the case in our study and whether the first impression has the same effect across our experimental conditions. As mentioned in Methodology, the order of the six statements are randomized for every crowd worker. Hence, we divide our participants into two groups: Good Impression (GI, the LLM's judgment is correct on the first statement) and Bad Impression (BI, the LLM's judgment is incorrect on the first statement). The number of workers belonging to group BI are 32 (Label), 35 (Explanation), 33 (Label+Exp). As shown in Figure 6, we look at the mean trust scores for the two groups with a breakdown over experimental conditions. We conduct a *Mann Whitney U test* to determine differences between *GI* and *BI* in each condition. We find that the mean trust scores for GI in condition Explanation is significantly higher than that for the BI group ($p = 0.0186$), indicating that crowd workers who have a good first impression report higher trust in the LLM. In contrast, we do not identify similar patterns in conditions Label and Label+Exp. This suggests that exclusively showing LLM-generated explanations has a stronger effect on crowd workers' 'first impression'.

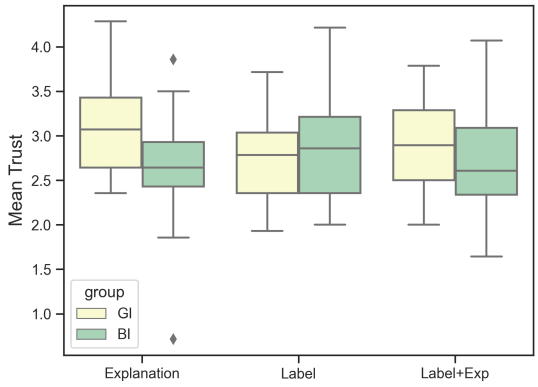


Figure 6: Distributions of mean trust score for the *BI* and *GI* groups over conditions.

One possible explanation is that crowd workers have the ability to validate LLM-generated explanations using their own knowledge or by using the customized search engine embedded in the task. As a result, they may choose to trust or disregard the LLM explanations based on their perceived statement credibility.

RQ4: Behavioral Indicators

Next we report on how LLM-generated answers impact workers' judging behaviors by looking at two indicators: (i) utilization of the search engine and (ii) assessment time. We propose two Null Hypotheses to test:

> **Null Hypothesis 5** *LLM-generated answers do not have an effect on the way that crowd workers interact with provided search engine.*
>
> **Null Hypothesis 6** *LLM-generated answers do not have an effect on assessment time.*

Utilization of the Search Engine. We first explore the interaction between crowd workers and the customized search engine provided with the assessment task and look at the impact of this interaction on their behaviour. Specifically, we

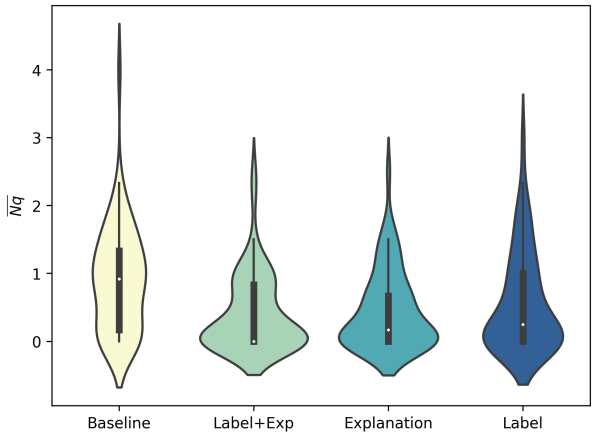


Figure 7: Average number of queries over experimental conditions.

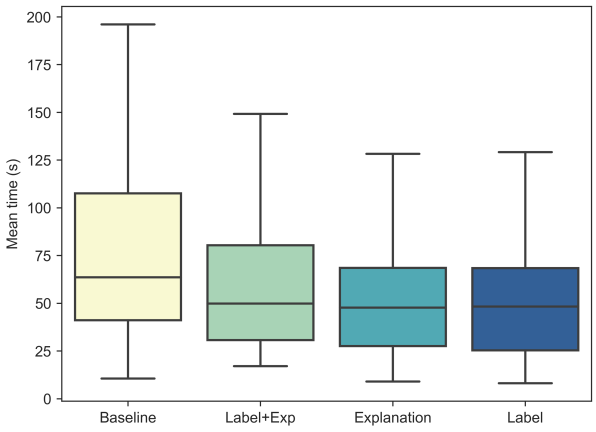


Figure 8: Average task completion time (s) across experimental conditions.

focus on the search queries made by crowd workers. Prior studies showed that making use of a search engine can be considered as an indicator of putting more effort in the task (Roitero et al. 2020).

We first look at the Number of Queries (Nq) throughout the task. Figure 7 shows the distribution of the number of queries issued by crowd workers across the four experimental conditions. Due to the non-normal distributions, we continue to apply ART ANOVA to investigate the effects caused by the presence of LLM-generated answers. The result suggests a significant interaction effect on $\overline{Nq}$ ($F = 6.5817, p = 0.0109$, partial $\eta^2 = 0.0271$). Subsequently, a pair-wise comparison (Tukey's HSD) reveals that crowd workers in condition Baseline issue significantly more queries as compared to workers in other conditions ($p < 0.01$ for all pairs). This can be a signal that crowd workers are taking the strategy of referring to the answers (labels and/or explanations) from GPT-3.5 to reduce the efforts towards retrieving relevant evidence to support their judgments (i.e., utilizing the search engine). Then, we take the median value of Nq for each condition as the pivot and, divide crowd workers into two groups: Search-Active Group (SA) and Search-Light Group (SL). By doing this, we are able to investigate the differences in performance between these two groups. We find that, for SA participants, there is no significant difference among the four experimental conditions in terms of ME_{LLM} (Kruskal-Wallis H Test, $H = 6.8693, p = 0.0762$), whereas in the condition Baseline SA participants have lower ME_{LLM} compared to those in the condition Label (Kruskal-Wallis H Test, $H = 11.6941, p = 0.0085$, post-hoc Dunn's Test $p = 0.0030$) and Label+Exp (post-hoc Dunn's Test $p = 0.0463$). As we have shown that labels in condition Baseline are significantly less aligned with GPT-3.5's labels as compared to the other conditions. Hence, we reckon that crowd workers are able to mitigate the bias from the LLM by leveraging the search engine results (i.e., group SA).

Assessment Time. Next, we focus on how much time crowd workers spent on the assessment task under different experimental conditions. Here, we utilize the *average time* taken across six statements as the metric to measure crowd worker effort. Figure 8 shows the average time used over experimental conditions. The median time spent under condition Baseline is higher than that in the other three conditions. We perform ART ANOVA to analyze the effects of LLM-generated labels and explanations towards task time. The interaction effect between the two factors is significant ($F = 5.5908, p = 0.0189$, partial $\eta^2 = 0.0239$). After conducting a post-hoc pair-wise comparison with Tukey's HSD adjustment, we observe that crowd workers in the Baseline condition spend significantly more time as compared to those in the Label condition ($p = 0.0371$). Considering the Label condition where only generated labels are available and the tendency of crowd workers to rely heavily on those labels, it is clear why they take less time to complete the task. On the other hand, crowd workers in Label condition are producing labels of comparable quality to those in the Baseline condition when looking at external agreement measures. This implies that solely providing LLM-generated labels can be an effective method to speed up crowdsourcing of misinformation assessment.

Discussion

RQ1: Quality. One notable observation regarding assessment quality is the significant overestimation errors made by crowd workers when LLM-generated information is presented. On the other hand, Crowd workers who are exposed to LLM-generated answers have fewer underestimation errors than those in the baseline condition. Given the similar tendency of LLM-generated labels leaning towards overestimating compared to the ground truth in the task corpus used in this study, it can be inferred that crowd workers' decisions are influenced by the information provided by the LLM. Although other studies addressing crowdsourced misinformation judgments have also reported the native overestimation errors made by crowd workers (Maddalena, Ceolin, and Mizzaro 2018; Roitero et al. 2020), our finding indi-

cates the impact of LLMs by looking at the judging quality across LLM-available (i.e. 'Label+Exp', 'Label' and 'Explanation') and no-LLM (i.e., 'Baseline') conditions. We also investigate judgment quality with respect to external and internal agreement. Statistical tests show that the judgment quality does not vary across conditions in terms of both types of agreement. Furthermore, we have observed similar levels of accuracy over the four conditions. In other words, crowd workers are mimicking the same overestimation errors done by the LLM causing their judgment quality to be not ideal for true statements. On the other hand, in the absence of support from the LLM, crowd workers make more underestimation errors and in turn, reach a similar level of judgment quality as LLM-supported workers. This shows how being exposed to LLM answers changes the labeling pattern of crowd workers.

RQ2: Confidence. Our analysis shows that the impact of answers from the LLM is minimal towards self-reported confidence level, as indicated by the similar distributions of confidence scores across the different experimental conditions. A linear regression shows a positive correlation between mean confidence and external agreement in the condition where both types of LLM outputs are presented. This suggests that in this particular condition, the LLM's outputs may reinforce pre-existing beliefs, leading to higher levels of confidence among crowd workers.

RQ3: Reliance and Trust. We investigate the extent to which crowd workers rely on the LLM. It is evident that crowd workers' labels have higher agreement with the labels given by the LLM, indicating a strong influence of the LLM. This influence can be classified as a bandwagon effect, similar to what has been observed in other crowdsourcing tasks involving advice from humans (Eickhoff 2018). This effect can be mitigated by concealing the truthfulness labels, as observed in the condition Explanation.

Solely showing LLM-generated explanations can also enhance confidence. Exclusively presenting explanations also triggers the effect of the 'first impression', which means that crowd workers express more trust in the model when the model initially provides correct information. This confirms the presence of reliance on LLM and the importance of LLM being accurate before being deployed as a co-pilot with human annotators.

RQ4: Behavior. To address RQ4, we investigated two behavioral indicators: the utilization of the custom Search Engine and task completion time. We observe a significant decrease in the use of the search engine when crowd workers are assisted by the LLM, suggesting that their working strategy relies on LLM answers to reduce the efforts they need to invest to complete the task. One piece of evidence that confirms this is that crowd workers in the condition Label spend significantly less time compared to those in the condition Baseline. We also notice that crowd workers who actively utilize the search engine exhibit lower reliance on the answers generated by the LLM. This suggests that in cases where the requester can disregard the potential over-reliance effect (Vasconcelos et al. 2023), they can consider displaying labels generated by an AI in order to expedite the process of manual misinformation assessment. These observations confirm previous results that show how crowd workers make use of additional tools and metadata to improve their work efficiency and, consequently, their hourly wage.

Implications

Our results have direct implications on how to embed LLM-generated answers in crowdsourced misinformation assessment tasks. Showing LLM-generated answers can be a double-edged sword. The presence of such answers does enhance the efficiency of crowd assessors, offering benefits to their hourly wage. However, this efficiency improvement is accompanied by possibly biased assessments as crowd workers tend to excessively rely on the AI's judgment. A potential approach to mitigate this over-reliance is to include explanations only, without the accompanying labels.

This study also raises some ethical concerns about how LLMs may be used to produce misinformation. Our results show that when the LLM offers correct explanations at the start, crowd workers report higher levels of trust in the system. This implies that by controlling the correctness of a few first pieces of LLM-generated information, individuals might perceive AI-generated information as reliable and this could be leveraged to propagate misinformation

Limitations

These are possible limitations for our study: i) Dataset: In this study, we only consider one dataset that is based on US politicians. This dataset may introduce inherent biases (e.g., political bias) in the decision-making process of US-based crowd workers; ii) Choice of LLM: In this study, we utilize GPT-3.5 to generate outputs. However, it is worth to explore alternative LLMs, considering that generative models may introduce biases and generate fictional content (Gallegos et al. 2023); iii) Interfance Design: In our study we show LLMs outputs in a static way. The way LLMs are integrated in the task may be done differently and can be more interactive such as providing feedback (Hettiachchi et al. 2021) and being a part of an iterative workflow (Little et al. 2010).

Conclusions

In this paper, we addressed research questions related to the impact of LLM output on human decision-making behavior when assessing the truthfulness of political statements. We conducted a large-scale user study by means of crowdsourcing measuring the impact of being exposed to the misinformation classification decisions generated by an LLM, together with a natural language explanation of such decisions. Our results show that crowd workers rely extensively on LLM decisions up to the point of making judgment errors due to misleading information generated by the LLM. Such over-reliance on the LLM is also shown in a reduced time needed to judge truthfulness and a reduced use of web search to find supporting evidence.

Acknowledgments

This work is partially supported by the Australian Research Council (ARC) Training Centre for Information Resilience (Grant No. IC200100022) and by the Swiss National Science Foundation (SNSF) under contract number CRSII5_205975.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Checco, A.; Roitero, K.; Maddalena, E.; Mizzaro, S.; and Demartini, G. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, 11–20.

Doroudi, S.; Kamar, E.; Brunskill, E.; and Horvitz, E. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2623–2634. ACM.

Draws, T.; La Barbera, D.; Soprano, M.; Roitero, K.; Ceolin, D.; Checco, A.; and Mizzaro, S. 2022. The effects of crowd worker biases in fact-checking tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2114–2124.

Duan, X.; Ho, C.-J.; and Yin, M. 2022. The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *Proceedings of the ACM Web Conference 2022*, 1685–1696.

Eickhoff, C. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of WSDM*, 162–170. ACM.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *ICWSM*, volume 16, 183–193.

Franke, T.; Attig, C.; and Wessel, D. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, 35(6): 456–467.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Gatto, J.; Basak, M.; and Preum, S. M. 2023. Scope of Pretrained Language Models for Detecting Conflicting Health Information. In *ICWSM*, volume 17, 221–232.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

He, G.; Kuiper, L.; and Gadiraju, U. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Hettiachchi, D.; Schaekermann, M.; McKinney, T. J.; and Lease, M. 2021. The challenge of variable effort crowdsourcing and how visible gold can help. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–26.

Heuer, H.; and Glassman, E. L. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–21.

Hoes, E.; Altay, S.; and Bermeo, J. 2023. Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims.

Jahanbakhsh, F.; Katsis, Y.; Wang, D.; Popa, L.; and Muller, M. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–27.

Körber, M. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, 13–30. Springer.

Krippendorff, K. 2011. Computing Krippendorff's alpha-reliability.

Kumar, S.; West, R.; and Leskovec, J. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, 591–602.

La Barbera, D.; Roitero, K.; Demartini, G.; Mizzaro, S.; and Spina, D. 2020. Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. In *42nd European Conference on IR Research, ECIR 2020, Part II*, 207–214. Springer.

Lekschas, F.; Ampanavos, S.; Siangliulue, P.; Pfister, H.; and Gajos, K. Z. 2021. Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*, 68–76.

Maddalena, E.; Ceolin, D.; and Mizzaro, S. 2018. Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing. In *CIKM Workshops*.

Qu, Y.; Roitero, K.; Barbera, D. L.; Spina, D.; Mizzaro, S.; and Demartini, G. 2022. Combining human and machine confidence in truthfulness assessment. *ACM Journal of Data and Information Quality*, 15(1): 1–17.

Roitero, K.; Soprano, M.; Fan, S.; Spina, D.; Mizzaro, S.; and Demartini, G. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *ACM SIGIR*, 439–448.

Roitero, K.; Soprano, M.; Portelli, B.; De Luise, M.; Spina, D.; Mea, V. D.; Serra, G.; Mizzaro, S.; and Demartini, G. 2021. Can the crowd judge truthfulness? A longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing*, 1–31.

Shu, K.; Wang, S.; and Liu, H. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 312–320.

Tolmeijer, S.; Christen, M.; Kandul, S.; Kneer, M.; and Bernstein, A. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–17.

Tolmeijer, S.; Gadiraju, U.; Ghantasala, R.; Gupta, A.; and Bernstein, A. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, 77–87.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Weinzierl, M.; Hopfer, S.; and Harabagiu, S. M. 2021. Misinformation adoption or rejection in the era of covid-19. In *ICWSM*, volume 15, 787–795.

Wobbrock, J. O.; Findlater, L.; Gergle, D.; and Higgins, J. J. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 143–146.

Wu, Y.; Jiang, A. Q.; Li, W.; Rabe, M.; Staats, C.; Jamnik, M.; and Szegedy, C. 2022. Autoformalization with large language models. *NeurIPS*, 35: 32353–32368.

Xu, F. F.; Alon, U.; Neubig, G.; and Hellendoorn, V. J. 2022a. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 1–10.

Xu, J.; Han, L.; Fan, S.; Sadiq, S.; and Demartini, G. 2022b. Does Evidence from Peers Help Crowd Workers in Assessing Truthfulness? In *Companion Proceedings of the Web Conference 2022*, 302–306.

Xu, J.; Han, L.; Sadiq, S.; and Demartini, G. 2023. On the role of human and machine metadata in relevance judgment tasks. *Information Processing & Management*, 60(2): 103177.

Xu, L.; Zhou, X.; and Gadiraju, U. 2020. How does team composition affect knowledge gain of users in collaborative web search? In *Proceedings of the 31st ACM conference on hypertext and social media*, 91–100.

Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; and Hashimoto, T. B. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Zhu, H.; Dow, S. P.; Kraut, R. E.; and Kittur, A. 2014. Reviewing versus doing: Learning and performance in crowd assessment. In *17th ACM CSCW Conference*, 1445–1455.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in the Limitation section
   (g) Did you discuss any potential misuse of your work? Yes, in the Limitation section
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see Dataset section
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? No, we did not include any new assets
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, it is an open-source dataset
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA, we are not curating or releasing new datasets.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? No, we have only included a screen shot as the full instruction would compromise anonymity.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see the Crowdsourcing Task Design section.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see the Crowdsourcing Task Design section.
   (d) Did you discuss how data is stored, shared, and deidentified? Yes