ChatGPT and Generative AI: A new era for crowdsourcing?

Evgenia Christoforou [1], **Jahna Otterbacher** [1,2]

[1]CYENS - Centre of Excellence, Nicosia, Cyprus
[2] Open University of Cyprus, Nicosia, Cyprus

Abstract

Generative AI systems, such as ChatGPT, have recently made their way into everyday life, setting off an alarm as to who uses them and how. Human computation via crowdsourcing has traditionally focused on problems requiring a "human touch," problems that machines cannot (yet) solve. This work explores how Generative AI affects the present and future of crowdwork. We have conducted a large-scale, light-weight survey of crowdworkers' activities and beliefs on three popular platforms (Amazon Mechanical Turk, Prolific and Clickworker), asking 1.400 crowdworkers located across three different continents for their input. Our results not only explore the use of Generative AI tools by crowdworkers in the completion of the task, but also, document the emergence of a new type of crowdsourcing task. Additionally, we found strong evidence that the attitude of crowdworkers towards Generative AI is associated with the platform in which they operate.

1 Introduction

OpenAI's ChatGPT[1] is clearly one of the most disruptive technologies in recent years. A large language model (LLM) built into an interactive chatbot, it was released to the public in November 2022, and within a month, it had become the "fastest growing consumer Internet app ever."[2] It took no time at all for the "alarms" to set off, with experts and non-experts alike speculating as to how it would transform our everyday lives. ChatGPT's influence on the labor market was perhaps one of the first issues of discussion, with analyses emerging as to which occupations are the most vulnerable (Zarifhonarvar 2023). Another obvious area of concern is that of education, and how it might transform the nature of student writing (Bishop 2023) such that radical new ways of instruction and evaluation are said to be urgently required (Tlili et al. 2023). Scientific research is one more area impacted, given that AI is increasingly able to "pass" as human, even in the case of generated abstracts read by scientists (Else 2023). In short, few technologies have caused such a "panic" across so many sectors, in such little time.

[1]https://openai.com/blog/chatgpt

[2]https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Systems such as ChatGPT represent a fundamental turning point in AI systems, due to their generative and open-ended nature. Unlike other forms of AI, such as algorithmic decision support, or search and recommendation systems, which produce closed-ended output such as decisions, ranked lists or descriptions, Generative AI systems are capable of producing new data, which is plausible, relative to that which it was trained on (Muller et al. 2022). In the case of ChatGPT, the system can generate new texts, which are fluent and often cannot be distinguished between those written by a human. This is arguably why the "alarm" has gone off with respect to how the nature of work will be impacted; even "elite" knowledge-based work, such as computer programming, can to an extent, be done by large language models, and in particular, ChatGPT (Castelvecchi 2022).

Of course, pretrained LLMs have been used for many years (Li et al. 2021), but they have only captured the attention of the public rather dramatically during the past year. This was the result of their being embedded into interactive chatbots, in contrast to their previous, behind-the-scenes roles in applications such as machine translation or search engine auto-complete. The interactivity of the chatbots highlighted their perceived "human level" linguistic abilities. The ChatGPT application in particular, has put this technology into an easy-to-use form, which anyone – including a crowdworker – can consult for a range of tasks (e.g., information searching, translation, stylistic change, etc.).

Furthermore, OpenAI is actively encouraging the use of ChatGPT for text analysis.[3] It provides guidelines and example prompts for creating text classifiers with its API, as well as a specific example for creating a sentiment detection classifier for Tweets. Not surprisingly, researchers working in applied natural language processing (NLP) have been quick to explore – and systematically evaluate – the use of ChatGPT in a range of text annotation tasks, such as Stance Detection, Hate Speech Detection, Sentiment Detection, Bot Detection, to name but a few (Zhu et al. 2023; Huang, Kwak, and An 2023). Such tasks were (and still are) often crowdsourced to allow researchers and practitioners to build high-quality datasets for machine learning (Vaughan 2017). However, lately, researchers are questioning whether ChatGPT

[3]https://platform.openai.com/docs/guides/completion/prompt-design

might be more reliable and consistent (Chmielewski and Kucker 2020), as well as economical (Wang et al. 2021) in labeling text, compared to crowdworkers.

Hence, AI chatbots – and Generative AI more broadly – represent a definitive point of change for crowdwork, from the perspective of task requesters, but also, for workers themselves. Crowdworkers are now faced with the natural question: "If AI can do my job, what is left for me to do?" As workers are the receivers of hundreds of available tasks everyday, including tasks that might now be a consequence of Generative AI, we survey their perspectives on the matter. Apart from assessing their views on how their work will be impacted by AI, we also asked them if they have performed a specific task closely associated with Generative AI.

In addition to ChatGPT, another AI system by OpenAI, DALL·E 2 [4] processes natural language in order to create realistic images and art. Stable Diffusion and Midjourney are yet two other AI systems which, like DALL·E can create photorealistic faces (Borji 2022), among other tasks. Thus, the need for human evaluation of understanding what is "real" (i.e., human-made) and what is "machine-made" is eminent. In this respect, in our study, we explicitly ask the crowdworkers if they have completed a crowdsourcing task that asked them to identify "machine-made" content, either text (e.g., generated by ChatGPT) or images.

Requesters on the other hand, are now faced with the question: "Can Generative AI perform as well as crowdworkers?". But even before considering this, another question should concern requesters, as well as the research community more broadly: "If everyone has access to Generative AI, so do crowdwokers. Is *human computation* still genuinely human?" Clearly, when a requester goes to a crowdsourcing platform, they are expecting to get human-generated responses. Otherwise, they would just go directly to ChatGPT, given all of the support / tools / prompts from OpenAI on how to use it. To our knowledge, while there is a surge of research evaluating the potential for Generative AI to replace crowdwork, there is no work yet asking the imminent questions: "Do crowdworkers (already) use AI chatbots to help them complete their crowdwork? Which changes to crowdwork do they anticipate as a result of these technologies?". To this end, we have surveyed crowdworkers across three continents and three crowdsourcing platforms.

1.1 Contribution

To our knowledge, this is the first study exploring the impact of Generative AI on the present and future of crowdwork, which puts the crowdworkers themselves at the center stage. The results of our large-scale survey over three popular crowdsourcing platforms (Amazon Mechanical Turk, Prolific and Clickworker), with 1.400 crowdworkers, across three different continents, aims to shed light on the following research questions:

RQ1. Do crowdworkers use AI chatbots to complete a task? Do they intend to use them in the future? On which type of crowdsourcing tasks?

RQ2. Are crowdworkers used as "detectors" of AI generated content?

RQ3. What are the views of the crowdworkers on their future, given the presence of Generative AI?

Analysis on the collected responses indicates that the nature of crowdwork is changing, with workers using AI chatbots to complete tasks, and with new types of crowdsourcing tasks on the rise. Our evidences points to human computation via crowdsourcing as being in a transitional phase, to which both task requesters and crowdworkers need to adapt.

2 Background and Motivation

2.1 Crowdsourcing and data for AI systems

Harnessing human intelligence, through the act of crowdsourcing, is one of the main data sources of Machine Learning (ML) (Roh, Heo, and Whang 2019), especially for label creation (Chang, Amershi, and Kamar 2017; Zhang, Wu, and Sheng 2016). Approaches like the ImageNet project (Russakovsky et al. 2015), led the way for many years, demonstrating the capabilities of crowdsouring in platforms like Amazons' Mechanical Turk (MTurk) [5] for dataset annotation. Other works in the area explored the capabilities of the crowd in content generation, like writing paragraphs of text (Salehi et al. 2017) or even more complex tasks like writing short fiction stories (Kim et al. 2017).

Naturally, data generation through crowdsourcing comes with well known limitation (Garcia-Molina et al. 2016) such as crowdworker incentivization (Ho et al. 2015), task abandonment (Han et al. 2019), aggregation of collected responses and ground truth inference (Zheng et al. 2017; Zhang, Wu, and Sheng 2016). One of the biggest issues though is data quality (Daniel et al. 2018; Perikleous et al. 2022) and reliability (Gadiraju et al. 2015) and many approaches have been proposed so far for improving data quality,such as increasing the quality of the reported labels (Barbosa and Chen 2019), or improving the design of the crowdsourcing task (Draws et al. 2021). The goal of this study is to explore the possibility that practitioners in the area of crowdsourcing will have to deal also with another form of unreliable crowdworkers, i.e., workers that partly or completely base their responses on information received by AI chatbots, like ChatGPT. Such behavior on the part of the crowdworkers, can possibly even intensify the known limitation of crowdsourcing recorded above, as new methods and techniques will need to be designed to account for this new "behavior" on the side of the crowdworkers. For example, resorting to crowdwork has been viewed as a valuable solution for building datasets representing diverse groups of people (Barbosa and Chen 2019; Aroyo and Welty 2015), but if every crowdworkers' source of information and aid in completing a task is now ChatGPT, will this still be the case?

Apart from data generation, crowdworkers are being used as part of hybrid crowd-machine intelligence systems complementing and expanding the capabilities of AI (Correia et al. 2023; Vaughan 2017; Rafner et al. 2021). As this is a relatively new field, additional considerations must be taken,

[4] https://openai.com/dall-e-2

[5] https://www.mturk.com/

since assuming that true human work will be returned by the crowd might not be the case and even worse, human work might be hard to be identified.

2.2 Crowdsourcing platforms

The philosophy of micro-task crowdsourcing platforms in general lines is the same: provide a service that connects humans eager to complete a task, according to a set of instructions, with task requesters in need of a humans' intelligence. Over the years, many crowdsourcing platforms have been build, but not all are made equal. Amazon's Mechanical Turk (MTurk) [6] has been the leading example of micro-task crowdsourcing platforms, with a vast amount of studies exploring the demographics of its crowd (Difallah, Filatova, and Ipeirotis 2018) and the reliability in terms of intentions (i.e., honest v. malicious v. spammers) (Gadiraju et al. 2015) and cognitive biases (Eickhoff 2018; Draws et al. 2021). The platform has disclosed that over 500,000 crowdworkers are available, but evidence show approximately 100,000 active crowdworker (Difallah, Filatova, and Ipeirotis 2018). Furthermore, the majority of crowdwokers appear to be resident in USA or India. Common crowdsourcing tasks include surveys, content creation, information finding, text annotations, audio and video transcriptions and visual tasks. The MTurk platform provides templates to requesters for creating a crowdsourcing task and hosting the task in the platform [7]. Regarding payment, the platform prompts the participants in rewarding crowdworkers after validating their crowdwork.

In contrast with MTurk, which is USA-based, Prolific [8] and Clickworker [9] are based in UK and Germany respectively. Thus, each platform has a different code of conduct. While MTurk and Clickworker are serving both the needs of the industry practitioners and researchers, Prolific has a service that is centered around online experiments (Palan and Schitter 2018), targeted at researchers especially in the fields of social and economic science. This makes it the only platform that provides a separate dataset with a large list of demographic information on each crowdworker that participated in the task. Contrary to the other two platforms, which provide solutions and templates for creating and hosting the crowdsourcing task at the platform, Prolific currently only allows for external study links to be posted to the platform by requesters. This set-up has particularly drawn a lot of survey studies to be posted at the platform, as they are easier to generate and maintain using one of the mainstream survey and experiment tools. Regarding the demographic of the crowd on Clickworker, no information on the number of active workers is available, but the platform claim that a size of about 4.5 million registered users [10], mainly located in US and Europe. Similarly, Prolific has a very large, USA, Europe and UK based crowd with over 130,000 participant [11]. Regarding payments, Prolific has a special mechanism for engaging with requester and prompting them to provide fair payments to crowdworkers. Payments in the Clickworker platform are automatic and not rewarding spammers is only feasible after providing proof of low quality work to the platform administrators. Finally, Clickworker is the only platform of the three, that in order for a task created by a requester to be posted in the platform in need to be inspected and approved by the platform administrators. Although similar in concept, the above mentioned differences among the three platforms make each environment unique and can partially dictate the behavior and opinions of crowdworkers either due to the nature of the task they come in contact or the way payments are designed and processed, or the code of conduct imposed by each platform. In this study, we will explore the crowdworkers' choice in seeking the help of AI chatbots in completing a crowdsourcing task and will explore the effect that the platform in which they operate can have on this decision. Additionally, we will explore how exposed are crowdworkers, in each of the three platforms, to tasks that use them to perform some form of a Turing test.

[6] https://www.mturk.com/

[7] https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI

[8] https://www.prolific.co/

[9] https://www.clickworker.com/

[10] https://www.clickworker.com/clickworker-crowd/

[11] https://www.prolific.co/prolific-vs-mturk

2.3 Generative AI and crowdsourcing

Until recently, crowdsourcing platforms were the go-to environment for harnessing creative work (Oppenlaender et al. 2020). With the emergence of Generative AI, researchers are exploring the possibility of replacing crowdworkers with AI chatbots, since they are a cheaper solution (Gilardi, Alizadeh, and Kubli 2023). However, yet another motivation is the perceived decline in the quality of data generated via MTurk (Chmielewski and Kucker 2020). A example is provided by a team of political scientists, who evaluated the ability of ChatGPT 3.5 to perform a set of annotation tasks on tweets conveying political content. Specifically, (Gilardi, Alizadeh, and Kubli 2023) used the prompt:

> "Here's the tweet I picked, please label it as [task]," where one of five tasks was used (relevance, topic detection, stance detection, general frame detection).

With trained political science students providing the gold standard labels, the researchers compared the performance of Master MTurkers and ChatGPT on the tasks. They found that even zero-shot classifications by ChatGPT were more accurate and consistent (i.e., high interjudge agreement) than crowdworkers, for all but the topic detection task.

Several recent works (Zhu et al. 2023; Kocoń et al. 2023; Hoes, Altay, and Bermeo 2023; Kuzman, Mozetic, and Ljubešic 2023) explored the capabilities of AI chatbots in a number of topics that traditionally required human intelligence, such as sentiment analysis, hate speech, genre detection, and disinformation detection. Evaluation of the quality of texts generated by algorithms is another area in which human computation has traditionally been used. That said, it is a difficult area in that there is little consensus as to how exactly humans should evaluate AI-generated texts (van der Lee et al. 2021). Wang et al. (Wang et al. 2023) conduced a preliminary study. They asked GPT to evaluate texts, using two different prompts – one asking it to rate the text on a numeric scale, and one using stars.

> Score the following [task-ins] with respect to [aspect] on a continuous scale from 0 to 100, where a score of zero means "[ant-aspect]" and score of one hundred means "perfect [aspect]". Note that [aspect] measures [aspect-ins]. [Conditioned Text] [Generated Text] Scores:
>
> Score the following [task-ins] with respect to [aspect] with one to five stars, where one star means "[ant-aspect]" and five stars means "perfect [aspect]". Note that [aspect] measures [aspect-ins]. [Conditioned Text] [Generated Text] Stars

In their experiment, they used five popular meta-evaluation datasets, containing human evaluations on a range of textual genres (e.g., summarization, data-to-text, story generation). The results showed that the AI-evaluations had strong correlations to those of humans, even on the story generation task, which is of a "creative" nature.

With respect to using ChatGPT in annotation tasks, Kocon refers to the AI as a "Jack of all Trades...but master to none" (Kocoń et al. 2023). They performed an extensive evaluation of its performance on 25 NLP-related tasks. Their conclusion is that it performs decently overall, but can be hit-or-miss on particular tasks (and/or in predicting certain classes) relative to bespoke state-of-the-art solutions. In a similar spirit, although they also found generally promising performance on the tasks they evaluated, (Zhu et al. 2023) were concerned that ChatGTP's accuracy is not uniform across classes. This also resonates with (Hoes, Altay, and Bermeo 2023) report on fact-checking, in which error rates were higher for debunking false information, as well as for older pieces of information.

Thus, it appears that ChatGPT capabilities are comparable, up to a certain degree, to human performance on annotation tasks. Since ChatGPT-3 is publicly available to everyone, if crowdworkers were to rely on this tool to help them complete their crowdwork, most probably requesters would have difficulties spotting "authentic" human work. To the best of our knowledge, ours is the first study asking the crowdworker the obvious question: Are you using ChatGPT to complete your crowdwork?

Apart from AI chatbots that have the capability of generating original text-based content, tools like DALL·E 2 [12], which generate images from text prompts are pushing the bounds on what is "real" or human[13], and what is "machine " generated. It appears that, with the rise of Generative AI, there is new and urgent need to understand under which circumstances humans can perceive what is "machine-made" (Shen et al. 2021). In this respect, our study poses another rather obvious but crucial question to the crowdworkers: "Were you ever asked, in a crowdsourcing task, to validate if some content or image was generated by Artificial Intelligence (AI) or a human?" Establishing the truthfulness of this statement can be fundamental for assessing the crowdworkers' perspective on what the future holds for them, and this is what we do by explicitly asking them this question.

[12]https://openai.com/dall-e-2

[13]We realize that this statement may be controversial, but leave further philosophical discussion out of the current work.

	Prolific	MTurk	Clickworker
USA	**199** (200)	**194** (200)	**183** (200)
UK	**99** (100)	-	**83** (100)
EU (GER, ITA, FRA, ESP)	**200** (200)	-	**189** (200)
India	-	**196** (200)	-

Table 1: Number of crowdworkers' responses (in parenthesis) and number of responses considered after cleaning for each country of residence and crowdsourcing platform.

3 Methodology

Our objective was to comprehensively document the present influence of generative AI on crowdwork. Thus, we created a straightforward survey that captures the following aspects: (1) experience of the crowdworker in performing crowdwork, (2) motivation of the crowdworker for performing crowdwork, (3) platforms used and types of tasks performed, (4) experience with ChatGPT, (5) use or future use of ChatGPT and how, (6) completion of crowdsourcing task for identifying human or machine generated context, (7) opinion of the crowdworkers on the impact of generative AI on crowdwork more generally.

We posted our survey to three distinct platforms: (1) Prolific, (2) MTurk and (3) Clickworker. On each platform, we aimed for a balanced number of responses gender-wise. Additionally, according to the platforms' declarations [14] [15] [16] and known crowd populations (Posch et al. 2022; Difallah, Filatova, and Ipeirotis 2018) we aimed at crowdworkers residing in a country with a known presence at each of the three platforms. Apart from specifying the country of residence criteria and guaranteeing a balanced gender sample, we did not impose any further restrictions on the crowdworkers that could complete our survey (e.g., master qualification on MTurk). Table 1 depicts the responses we collected from each platform in parentheses, and in bold, the final responses we considered after cleaning. We used an external link to administer our crowdsourcing task in order to provide the exact same survey across platforms; the cleaning process removed responses where the crowdworker failed to prove their identity in the platform (i.e., a valid crowdworker id). We collected in total 1.400 responses and after cleaning, we considered 1.343 gender-balanced responses in our analysis.

Crowdworkers were first informed of the goals and objectives of the study, as well as their right to withdraw from the survey at any point. The study received ethical approval from [redacted] and crowdworkers were rewarded fairly according to the respective platform's instructions, respecting

[14]https://researcher-help.prolific.co/hc/en-gb/articles/360009220833-Who-are-the-participants-on-Prolific-#heading-0

[15]https://www.clickworker.com/clickworker-crowd/

[16]https://www.mturk.com/help

	Prolific	MTurk	Clickworker
Single-platform users	65.9%	92.8%	61.8%
Experience:			
- Legacy	49%	17.2%	35.2%
- Newcomer	16.7%	3.8%	10.8%

Table 2: Participants platform demographics.

	USA	India	UK	EU
Prolific	44.7%	-	34.3%	57.5%
MTurk	94.3%	83.1%	-	-
Clickworker	48.6%	-	48.1%	43.4%

Table 3: Percentage of workers reporting use of AI chatbots, like ChatGPT, in everyday life, by platform and country.

the average hourly salary per country.

4 Data Analysis

4.1 Demographics of the Crowdworkers

Apart from collecting the country of residence and gender of the crowdworkrers who participated, we have also collected some additional information that describe the workers' crowdsourcing activity and experience. In particular, Table 2 presents some of the participants platform demographic we have collected and will be used in the analysis and discussion of the results. It is noteworthy, that we didn't ask participants for their level of experience, but we rather asked them to report on their years of experience with the platform and the amount of completed tasks. For the purposes of our analysis in this work, we define two categories of crowdworkers("newcomers" and "legacy"). "Newcomers" represent the extreme case of crowdworkers, with very few tasks ever completed ("between 0-10 tasks"), and with a short time-wise experience ("first time doing crowdsourcing"). On the other hand, "legacy" crowdworkers are the ones with more than 1 year of experience and more than 100 tasks completed. Within our questionnaire, we have collected separately the years of experience and the amount of completed tasks. We haven't based this parameter on any data collected from the platforms for the purpose of uniformity in our metric and consistency across platforms.

4.2 Crowdworkers and AI chatbots

One of the main goals of our study was to identify whether crowdworkers complete their tasks on their own, or if they are now seeking the help of Generative AI and specifically, ChatGPT, to complete a task. To establish that crowdworkers are aware of this disruptive new technology, we first posed the question of whether they have used AI chatbots to find information in their everyday life. Most MTurk crowdworkers (89.4%) reported having used an AI chatbot in their everyday lives, while that percentage drops to about half in Prolific (48.1%) and Clickworkers (48%). Table 3 presents the percentage of crowdworkers in each platform and country that use ChatGPT in their everyday life. For the responses received from Prolific, we perform a chi-square test of independence, to examine the relationship among the use of ChatGPT and the country of residence. Our null hypothesis is that the country of residence is independent of the choice to use ChatGPT in everyday life to collect information. It appears that for Prolific ($X^2(2, N = 498) = 15.48$, $p < 0.001$) and MTurk ($X^2(2, N = 390) = 12.14$, $p < 0.001$) there is a significant association among country of residence and use of ChatGPT while for Clickworker ($X^2(2, N = 455) = 1.16, p = 0.55$) there is not. Below, we explore the relationship of crowdworkers with AI chatbots.

Using ChatGPT to complete a task Table 4 provides an overview of the crowdworkers' behavior in the use of AI chatbots to complete a crowdsourcing task or to find information in their everyday life. The percentages presented considered the population of crowdworkers that declared that they are aware of this new technology. Less than 1% of the crowdworkers on Prolific and MTurk declared that they are unaware of AI chatbots and ChatGPT, while 3.2% on crowdworker on Clickworker declared that they were unware of this technology. It is apparent that crowdworkers, are aware of this new technology independently if they choose to use it or not. Results presented in Table 4 show that crowdworkers on the MTurk platform using AI chatbots (89.4%) have largely integrated the technology also on their crowdsourcing activities (77.8%). Only an 11.6% of those crowdworkers have opted from using it also to help them complete their crowdsourcing tasks. It is very interesting to observe that crowdworkers on the Prolific and Clickworker platform are behaving in an orthogonal manner. Apart from generally making use of this technology less ($\tilde{4}8\%$ of crowdworkers), only about half of the crowdworkers uses it in its crowdwork on Clickworkers, and a bit more than one third on Prolific. Interestingly, a very small percentage of crowdworkers, aware of the AI chatbot technology have claimed that they use it only for crowdwork and not in their everyday life activities. Below, we elaborate further on this finding and its implications.

From the collected results, it is clear the large infiltration of AI chatbot solutions to crowdwork on MTurk. In contrast, on Prolific and Clickworker crowdworkers do not seem to be eager to use this technology, at least in present time, to aid them in their crowdsourcing tasks. Our study, also explored the intentions of crowdworkers to use this technology in the future to aid them perform their crowdsourcing tasks. Considering the responses crowdworkers declaring that they are aware of the technology, an 86.8% of MTurk replied positively, a 23.8% on Prolific and a 36.4% on Clickworker replied positive. In all three platforms crowdworkers showed a larger willingness of using this technology in the future to help them perform crowdwork. In fact, crowdworkers' future intentions to use AI chatbots roughly doubled in size on Prolific and Clickworkers, but still remains well below MTurk.

Taking a closer look at the crowdworkers future intentions on the use of AI chatbot technology and the possible implications that this might have, we asked them to indicate to

ChatGPT Use \ Platforms	Prolific	MTurk	Clickworker
general use & crowdwork	10.9%	77.8%	17.5%
only crowdwork	2.2%	2.3%	3.4%
general use, no crowdwork	37.2%	11.6%	30.5%

Table 4: Use of AI chatbots, like ChatGPT, in everyday life and to aid in the completion of a crowdsourcing task, in each platform. (Percentages are calculated on the subset of crowdworkers who declared they are aware of the technology.)

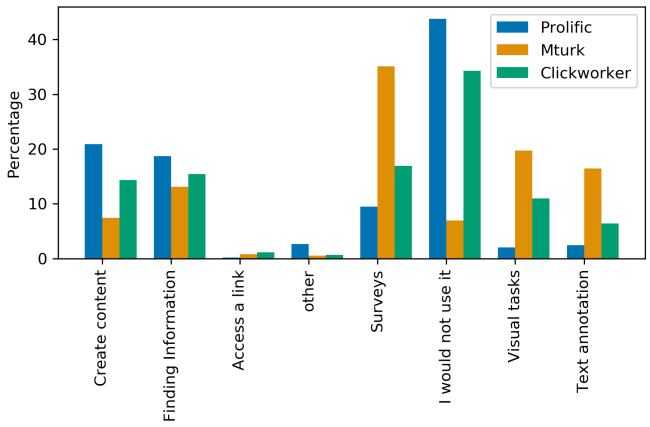

Figure 1: Responses: *For which type of task would you use an AI chatbot for?* (Closed form responses were provided in a more elaborated way to the participants.)

us for which type of crowdsourcing task they imagine using this technology for. Figure 1 clearly shows an unwillingness from Prolific and Clickworker crowdworkers to use an AI chatbot all together. MTurk crowdworkers seem to be more prone to use the ChatGPT technology to help them complete surveys. Prolific and Clickworker users seem to be thinking of using it to find information on an entity and create content, like translating a text, naming a product and writing a text. Responses received from Prolific and Clickworker participants seem like more obvious uses of ChatGPT to help them complete a crowdsourcing task, but in retrospect it is possible that MTurk participants find ChatGPT useful in the creation of large free text for open-ended survey questions. This is an interesting observation, that might hold severe implications on the reliability and "truthfulness" of future crowdsourced results. We provide a brief discussion on this later observation in Section 6.

Instructed to use AI chatbots As we have seen in Table 4, a very small percentage of crowdworkers have claimed that they use AI Chatbots only for crowdwork and not in their everyday life activities. After carefully inspecting these data, we noticed that a crowdworker on the Prolific platform, mentioned that the tasks she uses ChatGPT are: "One that requires I use it". This observation, has generated a new research question, one that we have not planned for in this study. Are crowdworkers instructed to use AI chatbots, like ChatGPT to complete crowdsourcing tasks?

Looking at the free text responses for the question (i.e., "other" in Figure 1): "For which type of task would you use an AI chatbot for?", Prolific participants gave us the following, enlightening free-text responses: "One that requires I use it", "I used it for a study that asked for ChatGPT responses to specific prompts", "When it is implemented as part of the task, when requested. I return true human work for all other tasks", "If the task in question explicitly asks for it as part of the task objective", "Tasks that asked to use AI assistance". Interestingly enough, only Prolific participants provided us with this type of responses, while participants from the other two platforms selected mainly the closed form-responses. In general, we notice that Prolific participants are more outspoken about their crowd activities. As we mention below, we plan to expand our study to better understand the possible implications in dataset creation when instructing crowdworkers to use AI chatbots to complete a task.

4.3 Crowdworkers as inspectors of AI generated content

Our study explored the connection of crowdwork with Generative AI, by asking crowdworkers whether they were asked to validate if some content or image was generated by AI or a human. It appears that the majority of crowdworkers have performed this type of task, with 92% of crowdworkers on MTurk, 50% on Prolific and 43% on Clickworkers, responding positively. Moreover, we have isolated the responses of crowdworkers who are "newcomers" (i.e., "first time doing crowdsourcing" and completing "between 0-10 tasks") and we observed that 87% has performed this task on MTurk, as compared to 42% on Prolific and a 29% on Clickworker. A chi-square test of independence was conducted to examine the relationship between platform and performing this task in the population of newcomers. The null hypothesis assumes that the platform is independent of whether a newcomer performed this task. The chi-square test ($X^2(2, N = 147) = 15,89$, $p < 0.001$) revealed a significant association . The above results, not only indicates that even a large percentage of "newcomers" experienced this type of task, but also, that there is an association with the platform they choose.

Looking at the "legacy" crowdworkers of a platform (i.e., "More than 1 year" crowdsourcing and completing "More than 100" tasks) we have noticed that 79% performed this type of task on MTurk, 56% on Prolific and 35% on Clickworker. So, this type of task is even more prominent among legacy workers, who are usually more sought out by requesters. The reason for a possible decrease in the MTurk

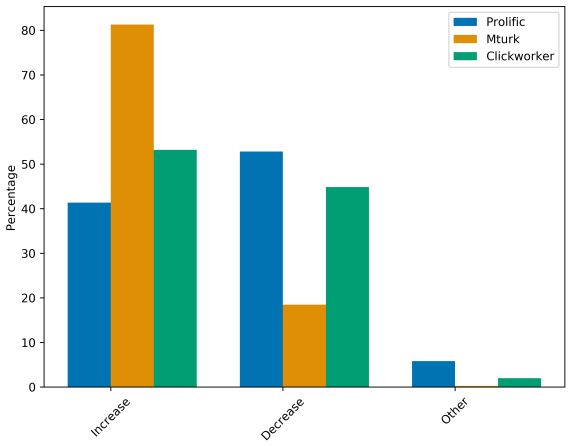

Figure 2: Responses: *AI chatbots or image generators will be responsible for increasing/decreasing the available tasks.*

platform is that experienced workers are priced higher and this could be a possible explanation for "legacy" having a lower percentage than "newcomers." Performing once more the chi-square test ($X^2(2, N = 471) = 40.62$, $p < 0.001$) reveals a significant association between the platform and a "legacy" crowdworker performing this type of task.

The above observations establish the conviction that tasks asking crowdworkers to inspect whether something is human-made or AI generated is a task currently present and quite active in the crowdsourcing platforms, since a large percentage of "newcomers" have completed it. Additionally, a large percentage of "legacy" (i.e., more experienced) crowdworkers have completed it. Thus, this type of task was completed by different "seniority" levels of crowdworkers, which indicates a sense of intensity and magnitude.

5 Crowdworkers' perspective on their future

On one hand, recent works explore the replacement of crowdworkers by Generative AI tools (Gilardi, Alizadeh, and Kubli 2023) and on the other hand, crowdworkers themselves experience the disruptive effect of Generative AI in one way or another. Our study harvested the opinion of crowdworkers both in an open and closed form question, since presently they are the ones experiencing in real time these changes. Figure 2 presents the responses across the three platforms on workers' beliefs of whether AI chatbots or image generators will be responsible for increasing/decreasing the available crowdsourcing tasks.

It is clear that crowdworkers on the MTurk platform have a more positive attitude on their future with the integration of Generative AI, while crowdworkers on Prolific have the most negative opinions. This observation is possibly associated with the fact that MTurk workers are embracing the use of ChatGPT and the large majority of them was also asked to identify if some content or image was created by an AI.

5.1 Analysis of Textual Responses

We performed an initial analysis of the open-ended responses of the workers, to the following question:

> Overall, do you think that Artificial Intelligence (AI) chatbots or image generators will have a positive impact on the practice of crowdsourcing? Please explain your thoughts on this.

Although we plan a more in-depth analysis in the near future, our current focus was on understanding any high-level differences in the quality of workers' responses, *across platforms*. To this end, we used the Linguistic Inquiry and Word Count (LIWC) gold standard software for analyzing word use (Boyd et al. 2022), focusing on two of its summary measures: Analytical Thinking and Emotional Tone.[17] The Analytical (Thinking) score represents an author's tendency to present more logical and/or formal expressions, while Emotional Tone quantifies the valence of the expression. They are standardized scores, which range from 1 to 99; thus, 50 represents an average/neutral score. Table 5 provides example responses that score high/low on these measures.

Many workers responded with only one word (e.g., yes/no) or even a short - but not very informative - response (e.g., "I don't know," "It's very helpful," "Yeah, I think so"). Therefore, to examine the quality of the responses received, we focused on those having more than the median number of words/response, in the given platform (Clickworker (9), MTurk (8), Prolific (13.5)). Table 6 details the median Analytic and Tone scores for responses at each platform, which are more extensive (i.e., longer) than the median response. As the scores were not normally distributed, we used a non-parametric Kruskal-Wallis test, to see if the there is a platform effect.[18] We followed with a pairwise Wilcoxon rank-sum test, to detect statistically significant differences.

As can be seen, the responses collected from workers at MTurk were generally more analytical and more positive in tone, than those collected from workers at the other two platforms. Thus, our analysis of the open-ended responses is very much in line with the findings of the analysis of the closed-ended responses. It appears that MTurkers share more analytical explanations on their views on the use of AI chatbots in crowdwork, as compared to the other workers.

5.2 Analytical Responses - A qualitative look

Finally, we explore several highly analytical responses (score > 90), which convey a clear negative, positive or neutral tone, to shed light on the workers' views on the prospects for AI chatbots to impact crowdwork.

Negative outlook. Some responses express a clear negative view on Generative AI, in that it will replace workers or at least reduce the number of available tasks. However, beyond that, some responses mentioned the lack of human

[17] Given the short, impersonal nature of the responses, we have not used Clout or Authenticity, which are more useful for analyzing texts in which the writer tries to establish a rapport with the reader.

[18] We use the following conventions to report p-values: **p<.001, *p<.01. P-values have been adjusted for multiple comparisons using a Bonferroni correction in all post-hoc tests.

Summary Measure	*High*	*Low*
Analytical Thinking	From the perspective of the workers, it's not a positive impact to have less jobs available.	I'm not sure, I don't have much insight in that topic.
	I think that for some tasks like image identification or labelling, AI chatbots can replace humans causing a decrease in the amount of available crowdsourcing tasks.	i think it will have a positive impact because adds more types of task and helps doing other types of tasks.
Emotional Tone	Yes I think it provides more opportunities with the need to train AI (at least initially).	I think it will have a more negative impact, because more jobs won't be available.
	Yes, because it's the future and will help human beings to improve and make their life easier.	I rather think it will have a negative impact because many crowdsourcing jobs (transcriptions, informations) can be done via AI.

Table 5: LIWC summary measures & example responses scoring high/low on each measure.

Platform	**# Responses**	**Analytic**	**Tone**
Clickworker (C)	219	26.10	20.23
MTurk (M)	194	76.97	83.95
Prolific (P)	249	30.77	20.23
Kruskal-Wallis Chi-squared		77.86***	36.92***
Post-hoc pairwise test		M > C,P ***	M > C,P ***

Table 6: LIWC comparison by platform - median scores.

experience and emotion, which requesters may aim to capture when crowdsourcing a task at a platform. Finally, another issue raised by Generative AI is that of transparency; while its use might create uncertainty as to who (or what) is performing the work.

1. No. It doesn't have a positive impact because it takes many freelancers life and also it don't give an real time data as compared to tasks done by human being. Most of the crowdsourcing tasks require a real time data and real time experience for the good result. As per my point of view, these AI or Image generators will be useful for checking purpose only. (MTurk - IN - M - More than 1 year - More than 100 tasks)
2. Artificial Intelligence (AI) chatbots or image generators give negative effect to crowdsourcing because they give ready made answered which is not unic and AI is not feel human emotion so some survey which want human emotion to solve problem so that case AI not work so i am not fever with AI they not give quality output. (MTurk - IN - F - More than 1 year - More than 100 tasks)
3. I believe the actual person should be doing the task to have more transparency in what is being accomplished in the crowdsourcing. (Clickworker - US - F - First timer)

Positive outlook. On the other hand, there are crowdworkers who express a generally positive outlook on the influence of Generative AI. For example, responses mentioned an increasing need for validation to be performed by crowdworkers, as well as the view that chatbots will never be able to replace genuine human responses. Still another view mentioned that the "AI arms race" means that crowdwork will constantly be required for testing new models.

1. Most of the crowdsourcing jobs I get are either surveys or validation tasks. I believe AI will be responsible for more tasks available for validation by a human worker. (MTurk - IN - M - More than 1 year - More than 100 tasks)
2. Probably positive, but I'm just guessing here. There will be more studies that deal with ALI chatbots for users to take. I don't see chatbots replacing users for surveys because people want real answers from humans and not just simulated answers from a bot. (MTurk - US - M - More than 1 year - More than 100 tasks)
3. They should have a positive impact because, it should be able to create more opportunities a computer and cannot fully predict the human mind. There will be more fact checking type assignments and constant testing of the AI. Some one will always come out with a new version. (Prolific - US - F - More than 1 year - More than 100 tasks)

Wait and see. A third type of responses expressed a "wait and see" view on the future of crowdsourcing with Generative AI. Some responses also expressed a view that we are currently witnessing some of the short-term changes, which may level off in the longer term.

1. In the short term, there may be a reduction in the available jobs - sure. But as AI improves and takes hold, new jobs may become available. I reckon we will have to wait and see how things pan out. (MTurk - IN - M - More than 1 year - More than 100 tasks)
2. I think it AI, will have an interesting effect, there are positives like being able to generate content faster, but also negatives like biases skewing results. We have to wait and see. (Prolific - US - M - Less than 6 months - More than 100 tasks)
3. I'm not really sure. It depends on whether the higher ups are fine with having people use AI for their work or if they just want to replace them. (Clickworker - US - F - More than 1 year - More than 100 tasks)

6 Discussion

The analysis of the collected responses clearly indicates that crowdwokers on MTurk choose to use AI chatbots to help them complete their crowdsourcing tasks (80.1%),

while only a small percentage of crowdworkers on Prolific (13.1%) and Clickworker (20.9%) resort to the help of AI chatbots. Regarding the crowdworkers' intentions to use this technology in the future, our respondents across all platforms showed an intention to integrate AI chatbots in their work. The above observations, partially answering RQ1, clearly indicate a "preference," or openness, of crowdworkers on the MTurk platform to use AI chatbots. As we have seen, this willingness to apply AI chatbots to their work can be attributed to the fact that almost all of them use the technology in their everyday lives (94.3%), as compared to crowdworkers in the other two platforms, where approximately only half of those in the collected sample reported using AI chatbots in their everyday lives.

Being in constant contact with AI chatbots, and observing their capabilities, it appears that crowdworkers, especially MTurkers, recognize that this technology can help them in completing a crowdsourcing task. When crowdworkers were asked to describe the type of tasks on which they would use an AI chatbot, more than 40% of workers on Prolific, and more than 30% of workers on Clickworker, categorically replied that they would not use it, with only about 7% of MTurkers providing the same answer. This observation, along with those made from the qualitative open responses (e.g., "... I return true human work for all other tasks," "I believe the actual person should be doing the task to have more transparency..."), indicates that many workers follow an unstated "moral code" in their crowdwork activities.

Regarding the second part of RQ1, it appears that surveys, content creation, information finding and text annnotation are among the most common tasks to which workers apply AI chatbots. We noticed that the second most common response of MTurkers was that of visual tasks for object identification, and this response comes rather as a surprise. A possible explanation for this result, is that crowdworkers resort to ChatGPT for tasks that require multiple labels for an object identified (e.g., using prompts such as "How might we describe an image of a flower?"). For instance, in tasks where workers are required to provide a number of different labels, ChatGPT might be consulted in coming up with additional descriptive terms.

Results collected from our study provide a positive answer to RQ2. It appears that crowdworkers are indeed used as "detectors" of AI-generated content. Furthermore, this type of task has been performed both by new and experienced crowdworkers. This indicates that crowdworkers, independently of their experience level, are already encountering this type of task, with MTurkers being more exposed to it, as compared to workers on the other two platforms.

Regarding RQ3, MTurkers have the most positive opinion of Generative AI and its influence on crowdwork, believing that this new technology will result in an increase in the available tasks. Analysis of the textual responses provided by the MTurkers also indicated a more positive tone when reporting their opinion on the future of crowdwork. More than half of crowdworkers at Clickworker have the same opinion as MTurkers, while workers in Prolific believe that Generative AI will impact the available tasks in a negative way (see Figure 2). A possible explanation for these platform differences, is the greater exposure of MTurkers both to AI chatbots, but also to tasks that ask them to detect AI-generated content. Thus, it appears that the above-mentioned stimuli may have led them to form a more positive opinion towards their future.

Our collected results provide clear evidence – particularly in the case of MTurk – that Generative AI has already disrupted crowdwork. One way this has happened is via new types of tasks, which ask workers to detect AI-generated content, or which ask them explicitly to use ChatGPT to execute the task (as described by some workers in their open-ended responses). Furthermore, crowdworkers reported that they use AI chatbots to help them complete a crowdsourcing tasks. Thus, many considerations arise relevant to data quality. For example, in the context of content generation, the issue of AI systems being self-alimented arises. AI chatbots become the creators and consumers of the same data (Aiyappa et al. 2023). Another example comes from image and text annotations where majority voting is a popular technique for data aggregation. If information received from crowdworkers becomes less diverse because they are all using the same AI chatbot, diversity in crowdsourced data will shift from being a competitive edge of the crowdsourcing approach to a utopia.

Given the above observations, new methods for improving the quality of the collected results are urgently needed. Although a number of detectors for flagging AI-generated texts has emerged recently (e.g., ZeroGPT[19], ChatGPT Detector[20], etc.), their effectiveness across a range of textual genres remains to be seen (Wang et al. 2023) and it may be the case that they introduce other problems (e.g., false positives on non-native speakers' writing (Liang et al. 2023)). We would venture to say that crowdworkers interacting with AI chatbots during their work does not necessarily invalidate it, but rather, the use of AI is something that must be transparent to the requester. Thus, it may be the case that platforms will need to establish a code of conduct on the use of AI in crowdwork.

6.1 Limitations and Future Work

We conducted a light-weight survey, asking a selected set of questions, aiming at a high engagement of the crowdworkers with our task. For this reason, we did not emphasize the type of crowdsourcing tasks workers would resort to with the help of AI chatbots, providing a short list of crowdsourcing tasks that are currently popular on commercial platforms. In the future, we plan to enrich our questionnaire with a more analytical list of crowdsourcing tasks, from which the worker can choose. Additionally, we plan to ask crowdworkers supplementary questions on how they would use an AI chatbot for completing a specific task. Finally, we aim at exploring the crowdworkers' point of view in having a code of conduct being established for the use of Generative AI in crowdwork.

[19] https://www.zerogpt.com

[20] https://chatgptdetector.co

7 Conclusions

To the best of our knowledge, this is the first study exploring the extent to which crowdworkers at various platforms use the help of AI chatbots for completing a crowdsourcing task. Our results clearly indicate that the large majority of crowdworekrs on MTurk use AI chatbots (e.g., ChatGPT) to help them complete a task, while on the contrary, the large majority of crowdworkers on Prolific and Clickworkers avoid the use of these technologies. Crowdworkers on MTurk and Clickworker have a generally positive attitude on what the future holds for them in light of the prevalence of Generative AI tools, while Prolific crowdworkers have serious concerns about their future in crowdsourcing. Furthermore, crowdworkers now receive tasks closely associated with Generative AI (e.g., detection and generation tasks). In summary, our results provide an idea of what the future of crowdwork might hold, but they also raise concern for the use of AI chatbots by crowdworkers, and what this might imply for the quality of the collected data.

Acknowledgments

This project has received funding from the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0421/0360 (KeepA(n)I), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578 (RISE), and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

Aiyappa, R.; An, J.; Kwak, H.; and Ahn, Y.-Y. 2023. Can we trust the evaluation on ChatGPT? *arXiv preprint arXiv:2303.12767*.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Barbosa, N. M.; and Chen, M. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Bishop, L. 2023. A computer wrote this paper: What chatgpt means for education, research, and writing. *Research, and Writing (January 26, 2023)*.

Borji, A. 2022. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47.

Castelvecchi, D. 2022. Are ChatGPT and AlphaCode going to replace programmers? *Nature*.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2334–2346.

Chmielewski, M.; and Kucker, S. C. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4): 464–473.

Correia, A.; Grover, A.; Schneider, D.; Pimentel, A. P.; Chaves, R.; De Almeida, M. A.; and Fonseca, B. 2023. Designing for Hybrid Intelligence: A Taxonomy and Survey of Crowd-Machine Interaction. *Applied Sciences*, 13(4): 2198.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1): 1–40.

Difallah, D.; Filatova, E.; and Ipeirotis, P. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 135–143.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170.

Else, H. 2023. Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944): 423–423.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 1631–1640.

Garcia-Molina, H.; Joglekar, M.; Marcus, A.; Parameswaran, A.; and Verroios, V. 2016. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4): 901–911.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Han, L.; Roitero, K.; Gadiraju, U.; Sarasua, C.; Checco, A.; Maddalena, E.; and Demartini, G. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 33(5): 2266–2279.

Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, 419–429.

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

Huang, F.; Kwak, H.; and An, J. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, 294–297.

Kim, J.; Sterman, S.; Cohen, A. A. B.; and Bernstein, M. S. 2017. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, 233–245.

Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Kuzman, T.; Mozetic, I.; and Ljubešic, N. 2023. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *ArXiv, abs/2303.03953*.

Li, J.; Tang, T.; Zhao, W. X.; and Wen, J.-R. 2021. Pretrained Language Models for Text Generation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track*.

Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; and Zou, J. 2023. GPT detectors are biased against non-native English writers. *Patterns*.

Muller, M.; Chilton, L. B.; Kantosalo, A.; Martin, C. P.; and Walsh, G. 2022. GenAICHI: Generative AI and HCI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7.

Oppenlaender, J.; Milland, K.; Visuri, A.; Ipeirotis, P.; and Hosio, S. 2020. Creativity on paid crowdsourcing platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Palan, S.; and Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17: 22–27.

Perikleous, P.; Kafkalias, A.; Theodosiou, Z.; Barlas, P.; Christoforou, E.; Otterbacher, J.; Demartini, G.; and Lanitis, A. 2022. How does the crowd impact the model? A tool for raising awareness of social bias in crowdsourced training data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4951–4954.

Posch, L.; Bleier, A.; Flöck, F.; Lechner, C. M.; Kinder-Kurlanda, K.; Helic, D.; and Strohmaier, M. 2022. Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics. *Human Computation*, 9(1): 22–57.

Rafner, J.; Gajdacz, M.; Kragh, G.; Hjorth, A.; Gander, A.; Palfi, B.; Berditchevskaia, A.; Grey, F.; Gal, K.; Segal, A.; et al. 2021. Revisiting citizen science through the lens of hybrid intelligence. *arXiv preprint arXiv:2104.14961*.

Roh, Y.; Heo, G.; and Whang, S. E. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4): 1328–1347.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Salehi, N.; Teevan, J.; Iqbal, S.; and Kamar, E. 2017. Communicating context to the crowd for complex writing tasks. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1890–1901.

Shen, B.; RichardWebster, B.; O'Toole, A.; Bowyer, K.; and Scheirer, W. J. 2021. A study of the human perception of synthetic faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8. IEEE.

Tlili, A.; Shehata, B.; Adarkwah, M. A.; Bozkurt, A.; Hickey, D. T.; Huang, R.; and Agyemang, B. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1): 15.

van der Lee, C.; Gatt, A.; van Miltenburg, E.; and Krahmer, E. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67: 101151.

Vaughan, J. W. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.*, 18(1): 7026–7071.

Wang, J.; Liang, Y.; Meng, F.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205.

Zarifhonarvar, A. 2023. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925*.

Zhang, J.; Wu, X.; and Sheng, V. S. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46: 543–576.

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

Zhu, Y.; Zhang, P.; Haq, E.-U.; Hui, P.; and Tyson, G. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.