🍄 Power-up! What Can Generative Models Do for Human Computation Workflows?

ANONYMOUS AUTHOR(S)

We are amidst an explosion of artificial intelligence research, particularly around large language models (LLMs). These models have a range of applications across domains like medicine, finance, commonsense knowledge graphs, and crowdsourcing. Investigation into LLMs as part of crowdsourcing workflows remains an under-explored space. The crowdsourcing research community has produced a body of work investigating workflows and methods for managing complex tasks using hybrid human-AI methods. Within crowdsourcing, the role of LLMs can be envisioned as akin to a cog in a larger wheel of workflows. From an empirical standpoint, little is currently understood about how LLMs can improve the effectiveness of crowdsourcing workflows and how such workflows can be evaluated. In this work, we present a vision for exploring this gap from the perspectives of various stakeholders involved in the crowdsourcing paradigm — the task requesters, crowd workers, platforms, and end-users. We identify junctures in typical crowdsourcing workflows at which the introduction of LLMs can play a beneficial role and propose means to augment existing design patterns for crowd work.

Additional Key Words and Phrases: crowdsourcing, generative AI, large language models, workflows, human computation

ACM Reference Format:
Anonymous Author(s). 2023. 🍄 Power-up! What Can Generative Models Do for Human Computation Workflows?. In *CHI'23: ACM CHI Conference on Human Factors in Computing Systems, April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

Artificial intelligence (AI) research is being reinvigorated with current advances in large language models (LLMs). Since their inception, LLMs have increased in size, effectiveness, and applications. For instance, BERT [10], initially trained for masked language prediction, has been applied to other domains such as neural ranking [18, 31] and document classification [2, 22]. OpenAI's[1] GPT family of models have been used in language tasks including goal-oriented dialogue [17], patent claim generation [23], and story generation [29]. The most recent GPT variant, ChatGPT [32], has seen an explosive growth in popularity, indicating the potential for a promising future where LLMs are deployed as work assistants. Due to such powerful generative capability, more researchers have started exploring generative LLMs in work assistant roles. For example, powerful generative LLMs have shown human-comparable writing skills in story generation [44] and scientific writing [16]. LLMs have also exhibited promising assistive capability in complex tasks like coding [13], drug discovery [28], and question generation for education needs [39].

The common thread running through all variations in LLMs is the need of high quality data for training and evaluation. Crowdsourcing has been widely adopted in machine learning practice to obtain high-quality annotations by relying on human intelligence [15, 37]. Crowdsourcing is a paradigm in which researchers or other stakeholders

[1] https://www.openai.com/

Manuscript submitted to ACM

request the participation of a distributed crowd of individuals, who can contribute with their knowledge, expertise, and experience [12]. Such individuals, called *crowd workers*, are asked to complete a variety of tasks in return for monetary or other forms of compensation. Tasks are often decomposed into smaller atomic units and can vary in their purpose, including labelling images, editing text, or finding information on specific topics [14]. Tasks can be standalone, or organized as a series of smaller sub-tasks, depending on their overall complexity and the design choices made by requesters. More complex problems, such as software engineering or system design problems, require task workflows.

Crowdsourcing workflows are distinct patterns that manage how large-scale problems are decomposed into smaller tasks to be completed by workers. The crowd-powered word processor Soylent applies the *Find-Fix-Verify* workflow to produce high-quality text by separating tasks into generating and reviewing text [6]. The *Iterate-and-Vote* workflow has been deployed in creating image descriptions, where workers are asked to write descriptions of images to assist those who are blind [26]. Subsequent voting tasks are used to decide on the optimal description. Chen et al. [9] introduce CrowdMR, which combines the *Map-Reduce* workflow with crowdsourcing to facilitate the solving of problems that require both human and machine intelligence, i.e., "AI-Hard" problems [42]. With CrowdForge, Kittur et al. [21] provide a framework for crowdsourcing to support complex and interdependent tasks. The authors follow up with the tool CrowdWeaver [20] for managing complex workflows, supporting such needs as data sharing between tasks and providing monitoring tools and real-time task adjustment capability. Taking a more holistic look at workflows, Retelny et al. [34] investigate the relationship between the need for adaptation and complex workflows within crowdsourcing, finding that the current state of crowdsourcing processes are inadequate for providing the necessary adaptation that complex workflows require.

Within crowdsourcing, the role of LLMs can be envisioned as akin to a cog in a larger workflow. Typically, LLMs are used for supporting individual writing or classification tasks within a workflow, as previous examples expressed. Researchers are also exploring the application of LLMs in assisting crowd workers. Liu et al. [27] combine the generative power of GPT-3 and the evaluative power of humans to create a new natural language inference dataset that produces more effective models when used as a training set. In a similar vein, Bartolo et al. [5] introduce a "Generative Annotation Assistant" to help in the production of dynamic adversarial data collection, significantly improving the rate of collection. These works measure the effectiveness of the models and the individual tasks, yet there remains an open gap regarding the understanding of how LLMs improve the effectiveness of crowdsourcing workflows and how such workflows can be evaluated.

In this work, we present a vision for exploring the gap from the stakeholders' perspectives, e.g., task requesters, crowd workers, and end-users. In so doing, we highlight the junctures of crowdsourcing workflows at which introducing LLMs can be beneficial. We also propose means to augment existing design patterns for crowd work.

2 INCORPORATING LARGE LANGUAGE MODELS IN CROWDSOURCING WORKFLOWS

As LLMs are pre-trained on large text corpora, they show great capability in understanding context-specific semantics. When further fine-tuned for specific uses with additional, smaller datasets, highly effective and domain-targeted models can be produced. Additionally, some LLMs (e.g., BART [24], GPT-3 [8]) are also good at generating responses to input queries, which can be fluent, human-like, and even professional. As it stands, LLMs have been effectively deployed within multiple domains such as medicine [4], finance [43], and others requiring commonsense reasoning [7]. As such, LLMs are an opportune and potentially very useful tool to use within crowdsourcing where domain knowledge may not always be available.

While LLMs are effective in many ways, they are far from being perfect and come with drawbacks. Due to their black box neural backbone, LLMs suffer from a lack of transparency, which leads to difficulty in explaining how they achieve the performances they do [45]. Such opacity also makes it difficult to track the factual error of LLMs, which inhibits the potential for improving the models [11]. Further, language models are known to capture several different forms of biases [1, 30, 38]. Most existing LLMs tend to perform poorly on tasks that require commonsense knowledge [33], which is a common practice for children. Last but not least, current language models achieve poor compositional generalization [19], which is required for solving complex tasks. Noticeably, LLMs fall short in aspects that humans are good at, e.g., commonsense reasoning [25] and complex task planning [3, 40]. Putting LLMs into practice requires either addressing or working within these limitations.

2.1 The Lens of Complex Crowdsourcing Workflows

LLMs can easily fit into existing crowdsourcing workflows. Take the Find-Fix-Verify workflow as an example. This workflow is well-suited for writing tasks, whether it be editing, revisions, or new content. Each step is an opportunity to include LLMs for improvements in the process. Let us take the example of revising a news story. During the "Find" stage, a workers would be tasked with reading the story and finding any errors, e.g., grammar, spelling, or false statements. Once these errors are identified, a new crowd of workers is recruited for the next stage: to "Fix" the errors. We are now left with an updated draft of the news story that has fewer errors than the initial draft. Which brings us to the final stage of the workflow, "Verify", where yet another group of workers validate the work of the prior groups. In this particular example, it is fairly clear where an LLM can be swapped for the workers at each stage. A retrieval or error classification LLM can be deployed for finding the errors, a generative LLM can be used to produce repaired text, and yet another classification LLM can finish it all off as the verifier. However, not all tasks take this form, or follow this particular workflow. Adapting other workflows, i.e., Iterate-and-Vote or MapReduce, can be done in a similar manner. Even so, adaptations such as these prompt the question: Once introduced, what are the effects of LLMs within crowdsourcing workflows for each stakeholder of the crowdsourcing process?

On the surface, this appears like straightforward question. Crowdsourcing has many different stakeholders involved: the *requesters*, the *workers*, and the *end-users*. The impact of including an LLM into workflows has the potential to affect each stakeholder in different ways. From the perspective of the requester, the monetary cost of completing tasks will be reduced as potentially fewer workers will need to be recruited. The tasks may take less time to complete which will result in further monetary savings. A reduction in time to gather data, complete tasks, and/or a reduced need for workers may have a negative impact on the income flow for workers, however. With available tasks taking less time and there being fewer tasks, it creates the potential for crowd workers to earn less. This can be offset by adjusting incentive structures on platforms. On the other hand, the reduction in costs for requesters could lead to more tasks being posted, leading to more high-quality labels. In turn, LLMs benefit from the better labels and improve in performance as well, creating a positive cycle that benefits both crowd workers and requesters. Further work is required to gain a better understanding of the financial opportunities and risks surrounding LLMs as part of crowdsourcing workflows.

Of course, there are trade-offs that come alongside any benefits. The trade-off for the requesters is a learning curve around the LLMs. Time will need to be dedicated to strategize and familiarize with the integration of LLMs in workflows. A trade-off that crowdsourcing platforms will share, accompanied by the additional cost of the development to add the LLMs to their products. An LLM must be trained before it can be appropriately used within a crowdsourcing workflow. This training, or fine-tuning, creates an overhead for either the crowdsourcing platform or the requester. While the overhead is initially a burden for most stakeholders, there will be an efficiency gain in the long term.

2.2 Risk and Opportunity

Further consideration is needed regarding the transparency of LLMs versus humans. When crowd workers complete tasks, such as annotation or other decision-oriented varieties, requesters have the capability of performing a follow-up with the workers to elicit reasoning for the outcomes provided. This is not a simple job for LLMs. While there exist methods for model explainability [35, 36, 41], none have demonstrated a level of effectiveness on par with what a requester would achieve with a human-human conversation. This same lack of transparency also has the potential of confounding workflows at the worker level. For example, take a scenario where an LLM is tasked with making a prediction, and a human worker to validate the prediction of the model, and the model provides a prediction that is not in line with what the worker expects to see. In such a scenario, the worker may want to interrogate the model to gain insight into why the prediction was made. However, there is currently no such clear way for the worker to request such an explanation from the LLM.

Also worth considering is the concept of accountability. Whenever a machine is introduced into a system, be it a factory, an airplane, or a crowdsourcing workflow, the question of accountability requires definition. Adding LLMs into crowdsourcing workflows raises the question of who or what is accountable if things do not go according to plan? Is the model, the requester, the platform, or the crowd workers to be held responsible for mishaps? There are many questions around the benefits, viability, risks, and harms involved with introducing LLMs into crowdsourcing workflows. These questions provide rich research opportunities for the generative AI and human computation research communities.

The realm of creative crowdsourcing tasks is another place of opportunity for LLMs. Generative models can help by providing suggestions or starting points to spark brainstorming or idea generation sessions. Alternatively, classification LLMs can be used to consolidate the ideas produced. For tasks that are more engineering or design focused, LLMs may be able to serve as "rubber duck" sounding boards. LLMs may also provide performance boosts in areas such as content creation, music composition, or protein discovery. The possibilities of how LLMs can be included in crowdsourcing are vast, yet the viability of these use cases warrants further investigation.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[2] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398* (2019).

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[4] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly Available Clinical BERT Embeddings. *NAACL HLT 2019* (2019), 72.

[5] Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2021. Models in the loop: Aiding crowdworkers with generative annotation assistants. *arXiv preprint arXiv:2112.09062* (2021).

[6] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 313–322.

[7] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317* (2019).

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Jun Chen, Chaokun Wang, and Yiyuan Bai. 2015. CrowdMR: Integrating crowdsourcing with MapReduce for AI-hard problems. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating Factual Knowledge in Pretrained Language Models. *arXiv preprint arXiv:2210.03329* (2022).

[12] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science* 38, 2 (2012), 189–200.

[13] Nat Friedman. 2021. Copilot: Your AI pair programmer. https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer/

[14] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*. 218–223.

[15] Ujwal Gadiraju and Jie Yang. 2020. What can crowd computing do for the next generation of AI systems?. In *2020 Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation*. CEUR, 7–13.

[16] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*. 1002–1019.

[17] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 583–592.

[18] Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-Rank with BERT in TF-Ranking. *arXiv preprint arXiv:2004.08476* (2020).

[19] Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, and Aaron Courville. 2022. On the Compositional Generalization Gap of In-Context Learning. *arXiv preprint arXiv:2211.08473* (2022).

[20] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver: visually managing complex crowd work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 1033–1036.

[21] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 43–52.

[22] Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical BERT with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems* 238 (2022), 107872.

[23] Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information* 62 (2020), 101983.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[25] Xiang Lorraine Li, Adhiguna Kuncoro, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2021. Do Language Models Learn Commonsense Knowledge? *arXiv preprint arXiv:2111.00607* (2021).

[26] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*. 68–76.

[27] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955* (2022).

[28] Zhichao Liu, Ruth A Roberts, Madhu Lal-Nag, Xi Chen, Ruili Huang, and Weida Tong. 2021. AI-based language models powering drug discovery and development. *Drug Discovery Today* 26, 11 (2021), 2593–2607.

[29] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*. 48–55.

[30] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[31] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[32] TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI* (2022).

[33] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4932–4942.

[34] Daniela Retelny, Michael S Bernstein, and Melissa A Valentine. 2017. No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–23.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

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

[38] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems* 33 (2020), 12388–12401.

[39] Xu Wang, Simin Fan, Jessica Houghton, and Lu Wang. 2022. Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs. *arXiv preprint arXiv:2205.00355* (2022).

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[41] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288* (2021).

[42] Roman V Yampolskiy. 2012. AI-complete, AI-hard, or AI-easy–classification of problems in AI. In *The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA*.

[43] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097* (2020).

[44] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[45] Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *Comput. Surveys* 55, 5 (2022), 1–31.