Communications of the ACM

Research and Advances

Artificial Intelligence and Machine Learning

Prevalence and Prevention of Large Language Model Use in Crowd Work

Crowd workers often use LLMs, but this can have a homogenizing effect on their output. How can we—and should we—prevent LLM use in crowd work?

By Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West

Posted Feb 19 2025

Crowd work platforms, such as Prolific and Amazon Mechanical Turk, play an important part in academia and industry, empowering the creation, annotation, and summarization of data,[11] as well as surveys and experiments.[21] At the same time, large language models (LLMs), such as ChatGPT, Gemini, and Claude, promise similar capabilities. They are remarkable data annotators[10] and can, in some cases, accurately simulate human behavior, enabling in-silico experiments and surveys that yield human-like results.[2] Yet, if crowd workers were to start using LLMs, this could threaten the validity of data generated using crowd work platforms. Sometimes, researchers seek to observe unaided human responses (even if LLMs could provide a good proxy), and LLMs still often fail to accurately simulate human behavior.[22] Further, LLM-generated data may degrade subsequent models trained on it.[23] Here, we investigate the extent to which crowd workers use LLMs in a text-production task and whether targeted mitigation strategies can prevent LLM use.

Key Insights

- LLMs are widely used in crowd work. We also find that responses written with the help of LLMs are high-quality but more homogeneous than those written without LLMs' help.
- LLM use by crowd workers compromises research on human behavior, preferences, and opinions. Our results indicate we must find ways to prevent inappropriate LLM use or appropriately incorporate LLMs into crowd workers' workflows.
- LLM use can be diminished by adding hurdles that complicate their use or by asking crowd workers not to use LLMs. However, LLM use remained common in our field experiments even with these mitigation strategies.

Study 1: Prevalence of LLM Use

To estimate LLM use on Prolific, a research-oriented crowd work platform, we asked n = 161 workers to summarize scientific abstracts (following Ribeiro et al.;[15] see Appendix in Supplemental Materials). We chose this task because it is laborious for humans but easily done by LLMs[17] and because it allowed us to use pre-LLM summaries from prior work[15] as "human ground truth." We ...ted whether a summary had been generated using LLMs with a fine-tuned e5-base classifier[28] ...ed on human, pre-LLM summaries[15] and summaries generated by GPT-4 and ChatGPT. The

model was then run on each of the 161 new summaries to estimate its probability of being LLM-generated. In this study, we did not instruct participants to use LLMs in any way; thus, we captured a baseline of LLM use for uninstructed participants doing a task for which LLMs have a considerable advantage over human labor.

Following a study on Mechanical Turk that took place four to six weeks before ours,[27] we used three approaches to aggregate the probabilities of LLM use (henceforth "LLM probabilities"), obtaining similar (but slightly lower) estimates:

- *Classify-and-count*, considering as synthetic any summary with an LLM probability above 50% (prevalence estimate: 33.3%; 95% CI [25.9%, 40.1%])
- *Probabilistic classify-and-count*, where we calibrated the model[6] (see Appendix) and then averaged the LLM probabilities (estimate: 35.2% [29.8%, 40.6%])
- *Corrected classify-and-count*, adjusting for the type I and type II error rates estimated on the training data[18] (estimate: 35.4% [27.8%, 43.0%]).

We validated our results by analyzing crowd workers' copy-pasting behavior (see Appendix), finding that 55% of the summaries where workers had copy-pasted text were classified as synthetic (that is, LLM probability above 50%) vs. only 9% when workers had not copy-pasted text. As no information about copy-pasting was used in the "LLM-or-not" classifier, this result strengthens our confidence in it. Interestingly, far fewer crowd workers used copy-pasting on Prolific (53%) in Study 1, compared with a previous study[27] on Amazon Mechanical Turk (89%).

Study 2: Prevention of LLM Use

Next, we analyzed whether targeted strategies can curb LLM use. Specifically, we studied two different mitigation approaches: 1) explicitly asking crowd workers not to use LLMs (henceforth the "request" strategy) and 2) imposing hurdles that deter LLM use (the "hurdle" strategy). We considered two variations for each: For the request strategy, we asked individuals either directly or indirectly not to use LLMs (see Appendix), and for the hurdle strategy, we either converted the original abstract text to an image or disabled copy-pasting entirely. As the two strategies are independent, we investigated all combinations (alongside a no-restriction condition) in a 3 x 3 factorial design (see Table 1).

Request	Hurdle: None	Hurdle: Image	Hurdle: Ctrl C+V
None	27.6% (21.0%, 34.6%)	21.5% (16.0%, 27.4%)	24.1% (18.3%, 30.4%)
Indirect	28.5% (21.7%, 35.8%)	19.8% (14.2%, 25.8%)	19.3% (14.6%, 24.5%)
Direct	24.0% (18.6%, 29.6%)	15.9% (11.9%, 20.3%)	15.8% (11.8%, 20.4%)

(a) Classifier

Request	Hurdle: None	Hurdle: Image	Hurdle: Ctrl C+V
None	15.8% (8.5%, 24.4%)	10.4% (3.9%, 16.9%)	4.9% (1.2%, 9.8%)
Indirect	13.2% (5.9%, 22.1%)	6.6% (1.3%, 12.0%)	3.6% (0.0%, 8.3%)
Direct	3.0% (0.0%, 7.1%)	6.6% (1.3%, 13.2%)	9.1% (3.9%, 15.6%)

(b) Self-reported

Request	Hurdle: None	Hurdle: Image	Hurdle: Ctrl C+V
None	10.9% (4.9%, 18.3%)	2.6% (0.0%, 6.5%)	1.2% (0.0%, 3.7%)
Indirect	4.4% (0.0%, 10.3%)	5.3% (1.3%, 10.7%)	2.4% (0.0%, 6.0%)
Direct	7.1% (3.0%, 12.1%)	4.0% (0.0%, 9.2%)	0.0% (0.0%, 0.0%)

(c) Heuristics

Table 1. LLM use across experimental conditions, estimated using three methods: a) probabilistic classify-and-count ("Classifier"); b) self-reported use ("Self-reported"); and c) high-precision heuristics ("Heuristics"). All estimates indicate that the interventions significantly reduced LLM use, albeit not completely.

Using the same task as in Study 1, Study 2 was conducted by randomly splitting n = 720 users into the nine conditions. Upon completion, they were then redirected to a follow-up survey where they were asked (Q1) how often they used ChatGPT in their daily lives, (Q2) whether they had used ChatGPT for the task, and (Q3) whether they knew of studies tracking ChatGPT use on crowd work platforms (see Appendix for exact phrasing). We measure LLM use with the probabilistic classify-and-count classifier, self-reported LLM use as captured by Q2, and high-precision (and likely low-recall) heuristics indicating LLM use (see Materials and Methods).

Effectiveness of preventive measures. Table 1 shows the estimated LLM use across different mitigation strategies. For example, when workers were directly requested not to use LLMs and shown the text to be summarized as an image (thus preventing copy-pasting), LLM use (as measured by the probabilistic classify-and-count method) almost halved, dropping from 27.6% to 15.9% (as measured by the classifier; see Table1a). Similar results were obtained using self-reported use by crowd workers (Q2) and using high-precision heuristics (Tables 1c and 2c; see Materials and Methods). Comparing high-precision heuristics with self-reports revealed that only 11 of the 31 workers using LLMs according to high-precision heuristics admitted to using LLMs, whereas 31 of the 689 whom the heuristic and classifier both failed to mark as synthetic admitted to LLM use.

We further disentangled the effect of each specific strategy and variation with a linear model (see Appendix), finding three out of the four tested interventions to significantly reduce LLM use (considering the LLM use predicted by the classifier; see the figure). Notably, asking crowd workers indirectly ("Please do your best to summarize the abstract in your own words") was the least effective strategy across all measures of LLM use and the only non-significant intervention when considering the classifier-based outcome ("Indirect"; 2% decrease; p = 0.38). This hints at the complexity of preventing LLM use, as crowd workers may choose to ignore requests if it is in their best interest financially.

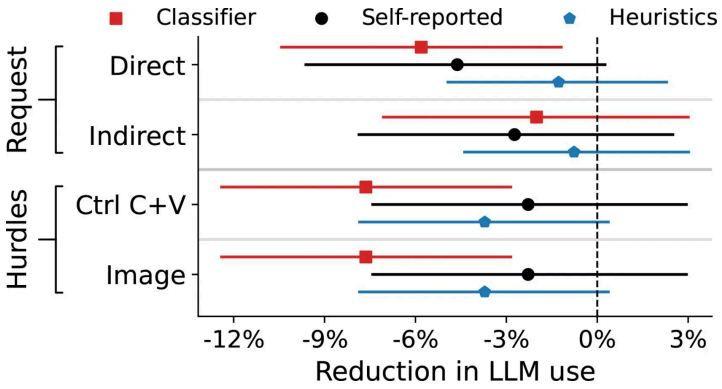

Figure. Estimated effect sizes for interventions to prevent LLM use considering three different measures of LLM use as the outcome variable: 1) probabilistic classify-and-count, 2) self-reported use, 3) high-precision heuristics. Error bars represent 95% confidence intervals; n = 720.

Correlates of LLM use. We studied the relationship between LLM use and 1) the age of crowd workers and 2) how they answered two of the post-survey questions (Q1: LLM use in general; Q3: awareness of studies measuring LLM use) using a simple linear model and considering both self-reports and the classifier's LLM-probability estimates as outcomes (see Appendix). We found that younger individuals were significantly more likely to use LLMs (−0.18% in estimated LLM probability per year; p = 0.014) and that workers who used LLMs "often" were 18.7% more likely to use it for the task (p < 0.001). Awareness of studies measuring LLM use did not significantly affect use (+1.6%; p = 0.55). Results were similar when considering self-reported use as the outcome variable.

Additionally, we analyzed the relationship between LLM use and time spent on the task, finding that preventive measures (that is, hurdles and requests) seem to mediate the relationship. Users who self-reported LLM use spent 21.9% less time (relative decrease; p = 0.002) to complete the task than

those who did not (across experimental conditions). Using a simple linear model with time spent as the log-transformed outcome (see Appendix), we further analyzed this relative change across different proxy metrics for LLM use and experimental conditions (see Table 2). Across proxy metrics, we found that the overall time reduction is never statistically significant when hurdles are employed. However, when only requests are applied, results differed: The relative decreases were not statistically significant considering the classifier but remained statistically significant considering self-reports. We hypothesize this may be because users who use LLMs and lightly edit their output spend more time on the task and are less likely to self-report use.

		Hurdle		
		None	Image	Ctrl C+V
Request	None	-32.0% (-51.6%, -4.5%)	18.0% (-19.8%, 73.6%)	5.1% (-22.6%, 42.7%)
	Indirect	-2.0% (-28.8%, 34.9%)	42.4% (-3.3%, 109.6%)	4.9% (-26.8%, 50.2%)
	Direct	-12.0% (-34.2%, 17.7%)	1.0% (-43.9%, 81.7%)	35.1% (-6.9%, 96.1%)

(a) Classifier

		Hurdle		
		None	Image	Ctrl C+V
Request	None	-34.2% (-55.3%, -3.2%)	-0.1% (-34.8%, 53.2%)	1.4% (-41.5%, 75.7%)
	Indirect	-43.8% (-61.0%, -19.1%)	-25.1% (-58.0%, 33.3%)	-17.6% (-56.0%, 54.1%)
	Direct	-58.9% (-78.5%, -21.5%)	1.8% (-43.4%, 83.2%)	22.7% (-15.7%, 78.8%)

(b) Self-reported

		Hurdle		
		None	Image	Ctrl C+V
Request	None	-46.1% (-65.4%, -16.1%)	35.3% (-40.2%, 206.3%)	-45.6% (-81.3%, 58.8%)
	Indirect	-53.7% (-75.0%, -14.3%)	-47.9% (-72.2%, -2.5%)	45.1% (-32.1%, 210.1%)
	Direct	-32.3% (-56.5%, 5.2%)	29.4% (-38.6%, 172.6%)	0.0% (0.0%, 0.0%)

(c) Heuristics

Table 2. Relative differences in time spent between instances where we detected LLM use and where we did not. We report differences across the nine experimental conditions and determine LLM use using three methods. (Note that time spent is one of our heuristics for detecting ChatGPT use.)

Content-level analysis. Analyzing the text of crowd workers' summaries, we found that summaries labeled as synthetic by the classifier were significantly more "homogeneous" than those labeled as human, according to a previously proposed homogeneity metric[20] and BertScore[30] (details in Appendix). We estimated a homogeneity score of 45.6% (43.2%, 48.2%) for synthetic texts, vs. 27.1% (26.8%, 27.4%) for human texts, and a BERTScore of 91.4 (91.0, 91.8) for synthetic texts vs. 87.4 (87.2, 87.3) for human texts.

In the original study whose human summaries we reused,[15] the authors measured the retention of keywords from the original abstract corresponding to essential information, finding it to be highly correlated with human evaluations of quality. Using this metric as a proxy for quality, we found that summaries labeled as synthetic preserved more keywords (40.1% [36.9%, 43.2%]) than summaries labeled as human (31.2% [29.9%, 32.6%]). We found a similar effect when using self-reports and high-precision heuristics instead of the classifier's labels.

But how do the interventions affect the above content-level metrics? We repeated the analysis shown in the figure but using homogeneity, BERTScore, and keyword retention as outcomes (see Section G.1 in the Appendix for details). We found no significant effect of the interventions on content-level outcomes, with one exception: Directly requesting workers not to use LLMs decreased keyword retention by 5.8% ($p = 0.003$). We hypothesize that the reduction in keyword retention may be caused by crowd workers' hesitancy to use extractive summarization when prompted not to use LLMs. (Results were similar when considering only summaries classified by us as being human-made.)

Discussion

The results suggest that LLMs pervade current crowd work on text-production tasks. Although adopting various strict mitigation approaches reduced LLM use by nearly 50%, it could not entirely prevent it. While text-production tasks are particularly suitable for LLM use, we argue that these findings are broadly applicable to crowdsourcing, as crowd workers will likely use LLMs on other kinds of tasks (for example, image segmentation, multiple-choice questions) in the near future, if they are not already doing so. There are several reasons for this. First, the models are increasingly capable of doing other tasks; for instance, while writing this article, ChatGPT was updated to receive images as input,[19] which could allow its use on tasks such as image tagging or classification.[29] Second, crowd workers have incentives to use them; even in the absence of LLMs, there are widespread attempts to "game the system" to make money, to the extent that an extensive body of work has been developed around ensuring the quality of responses.[7] Third, crowd workers, who are are often tech-savvy[5,12] and frequently rely on plug-ins and Web services to boost their performance and earnings,[9,16] are capable of integrating these models into their pipelines. Even without requiring coding, tools to automate ChatGPT use are plentiful (for example, IFTTT, Zapier).

Synthetic data may harm the utility of crowd work platforms, as researchers often care about human behavior or preferences; for example, the authors of the paper whose human summaries we borrowed[15] wanted to know *how people summarized*, instead of merely obtaining good summaries. While some preliminary studies suggest that synthetic data may capture certain viewpoints,[2] it still often fails to do so, and research using crowd work may inadvertently capture the behavior and preferences of LLMs, not humans. Even if LLMs can capture average behavior or preferences, the homogeneity of their responses may result in losing the long tail of human behavior and preferences that is vital to researchers[24] and, according to recent work, important to training capable LLMs.[23] In that context, our results indicating that LLM-generated summaries are more homogeneous than human-generated summaries suggest that LLM use may be particularly harmful when the goal of crowdsourcing is to capture the diversity of human preferences, behaviors, or opinions.

To foresee the potential harm of LLM use in crowdsourcing, one may consider a topic that has received increased attention in the social sciences in the past few years: climate change.[8,26] Social scientists often use crowdsourcing to study attitudes toward climate change.[4,25] Yet, recent work has shown that, when prompted to answer multiple-choice questions, LLMs' opinions are better aligned with liberal, wealthy individuals and exhibit pro-environmental bias.[13,22] Therefore, it may be expected that LLM use could harm the validity of social scientists' studies on behavioral interventions and assessments of global stances toward climate change. We stress that this is not particular to climate change: Social scientists use crowdsourcing for various topics,[3] and LLMs are non-representative of the samples of interest in various ways.[13,22]

We must be careful not to conflate LLM use with cheating. Depending on the study, it could be beneficial if LLMs assist crowd workers. Further, as LLMs become intertwined with how people write and accomplish everyday tasks, the distinction between "synthetic" and "human" data may blur. For example, is text generated with the help of a spellchecker synthetic? Thus, we expect the thresholds for concern and meaning will shift dramatically over the coming months and years, as LLMs become more ubiquitous in everyday productivity tasks. In that context, a fruitful future direction is to explore the landscape of how crowd workers use LLMs. There are many ways of integrating these models into crowd workers' workflows, and different approaches may affect downstream research output differently.

We found that stricter mitigation approaches can significantly reduce LLM use. These measures may, however, backfire when detection is critical. Stricter measures may limit the number of participants using LLMs but also make them more reluctant to admit ex-post that they used them, or make them more difficult to detect, as the prevention measure eliminated a key indicator of LLM use. For example, removing copy-pasting makes it harder to use LLMs, limiting use, but then researchers also cannot use copy-pasting as a feature to detect who used LLMs. Further, mitigation approaches can reduce the overall response quality: As we found empirically, workers explicitly told not to use LLMs produced lower-quality summaries.

LLM-based tools and LLM users are co-evolving in ways to ensure the low temporal validity of our specific findings and estimates. In the past few months alone, tools have evolved to interpret images

and to call LLMs without the need to copy-paste (for example, by simply selecting text). This does not diminish the value of our work—it makes it even more valuable: It is critical to establish baselines and ongoing measurements as this co-evolution progresses, and our work establishes such baselines. Further, we are confident that our high-level interpretations and guidance will translate across this evolution, and we hope this helps establish a regularly updated new program of study to serve crowd work platforms and researchers.

To conclude, in light of our findings, we propose practical guidelines for researchers to use crowdsourcing in the era of large language models. First, researchers should assess the impact of LLMs on their research by asking themselves: Is the point of crowdsourcing to obtain data representative of human behavior, preferences, and opinions? And if so, is capturing the diversity of these human responses important? We argue that crowdsourcing will be most affected when the answer to both questions is yes, as we found that LLM responses differ from human responses and are more homogeneous. Second, if large language models are likely to harm the utility of crowdsourcing, our findings indicate that researchers can actively diminish LLM use by requesting that workers not use them and creating hurdles that decrease the incentives for using them. Notably, hurdles should be adapted as models become more capable and better integrated into people's lives.

Materials and Methods

- **Data.** We modified a prior Mechanical Turk task[15] where crowd workers were asked to summarize medical paper abstracts. We re-ran the study twice on Prolific. In Study 1, we estimated prevalence by collecting 168 user summaries (paying £9/hour). In Study 2, we re-ran the study on 720 users, now using several mitigation techniques (paying £10/hour). (See Appendix for full description of data and original study.)
- **Model training.** We fine-tuned a e5-base-v2 language model[28] for our classification task and conducted a hyperparameter sweep. The model was trained on the summaries from the original study[15] (written before the adoption of LLMs) and summaries synthetically generated using OpenAI's API.
- **Heuristic-based estimates.** We defined two high-precision heuristics for measuring LLM use: feasible time for completion and pasting in artifacts from the ChatGPT Web interface (details in Appendix).
- **Effect of each intervention.** We assessed the effectiveness of each of the interventions with a linear probability model. We do not consider interactions between the treatment conditions, as a two-way ANOVA indicated that the interactions between the two strategies are not statistically significant.

To access the Appendix, please visit this article's page in the ACM Digital Library and click on Supplemental Material.

References

References

1. Akiba, T. et al. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining*. ACM, (2019), 2623–2631.

2. Argyle, L.P. et al. Out of one, many: Using language models to simulate human samples. *Political Analysis 31*, 3 (2023), 337–351.

3. Bohannon, J. Mechanical Turk upends social sciences. *Science 352*, 6291 (2016).

4. Bouman, T., Steg, L., and Zawadzki, S.J. The value of what others value: When perceived biospheric group values influence individuals' pro-environmental engagement. *J. of Environmental Psychology 71*, 101470 (2020).

5. Brewer, R., Morris, M.R., and Piper, A.M. Why would anybody do this?: Older adults' understanding of and experiences with crowd work. In *Proceedings of the 2016 CHI Conf. on Human Factors in Computing Systems*. ACM, (2016), 2246–2257.

6. Card, D. and Smith, N.A. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, (2018), 1636–1646.

7. Daniel, F. et al. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys 51*, 1 (2018), 1–40.

8. Dietz, T., Shwom, R.L., and Whitley, C.T. Climate change and society. *Annual Rev. of Sociology 46* (2020), 135–158.

9. El Maarry, K., Milland, K., and Balke, W-T. A fair share of the work? The evolving ecosystem of crowd workers. In *Proceedings of the 10th ACM Conf. on Web Science*. ACM, (2018), 145–152.

10. Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. In *Proceedings of the National Academy of Sciences of the United States of America 120*, (2023).

11. Gray, M.L. and Suri, S. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, (2019).

12. Guess, A.M. and Munger, K. Digital literacy and online political behavior. *Political Science Research and Methods 11*, 1 (2023), 110–128.

13. Hartmann, J., Schwenzow, J., and Witte, M. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768* , (2023).

14. Hausman, J.A., Abrevaya, J., and Scott-Morton, F.M. Misclassification of the dependent variable in a discrete-response setting. *J. of Econometrics 87*, 2 (1998), 239–269.

15. Ribeiro, M.H., Gligoric, K., and West, R. Message distortion in information cascades. In *The World Wide Web Conf.* ACM, (2019), 681–692.

16. Irani, L.C. and Silberman, M.S. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, (2013), 611–620.

17. Luo, Z., Xie, Q., and Ananiadou, S. ChatGPT as a factual inconsistency evaluator for text summarization. *arXiv:2303.15621* , (2023).

18. Meyer, B.D. and Mittag, N. Misclassification in binary choice models. *J. of Econometrics 200*, 2 (2017), 295–311.

19. OpenAI. ChatGPT can now see, hear, and speak. (Sep. 25, 2023); https://openai.com/index/chatgpt-can-now-see-hear-and-speak/

20. Padmakumar, V., He, H. Does writing with language models reduce content diversity? In *Proceedings of the 12th Intern. Conf. on Learning Representations*. IEEE, (2024).

21. Salganik, M.J. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, (2019).

22. Santurkar, S. et al. Whose opinions do language models reflect? In *Proceedings of the 40th Intern. Conf. on Machine Learning.* PMLR, (2023).

23. Shumailov, I. et al. Model dementia: Generated data makes models forget. *Nature 631* (2024), 755–759; 10.1038/s41586-024-07566-y

24. Song, Z. et al. *Reward collapse in aligning large language models. arXiv preprint arXiv:2305.17608* , (2023).

25. Sparks, A.C. Climate change in your backyard: When climate is proximate, people become activists. *Frontiers in Political Science 3*, 666978 (2021).

26. Steg, L. Psychology of climate change. *Annual Rev. of Psychology 74* (2023), 391–421.

27. Veselovsky, V., Ribeiro, M.H., and West, R. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, (2023).

28. Wang, L. et al. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, (2022).

29. Yang, Z. et al. The dawn of LLMs: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, (2023).

30. Zhang, T. et al. BERTscore: Evaluating text generation with BERT. In *Intern. Conf. on Learning Representations*, (2020).

About the Authors

Veniamin Veselovsky is a doctoral student in the Department of Computer Science at Princeton University.

Manoel Horta Ribeiro (manoel@cs.princeton.edu) is an assistant professor in the Department of Computer Science at Princeton University.

Philip J. Cozzolino is an associate professor in psychiatry and neurobehavioral sciences at the University of Virginia School of Medicine.

Andrew Gordon is a staff researcher in social and behavioral science at Prolific.

David Rothschild is an economist at Microsoft Research.

Robert West is an associate professor in the School of Computer and Communication Sciences at EPFL and a visiting researcher at Microsoft Research.

Join the Discussion (0)