Laying foundations to quantify the "Effort of Reproducibility"

Akhil Pandey Akella
Northern Illinois University
Dekalb, Illinois, USA
aakella@niu.edu

David Koop
Northern Illinois University
Dekalb, Illinois, USA
dakoop@niu.edu

Hamed Alhoori
Northern Illinois University
Dekalb, Illinois, USA
alhoori@niu.edu

ABSTRACT

Why are some research studies easy to reproduce while others are difficult? Casting doubt on the accuracy of scientific work is not fruitful, especially when an individual researcher cannot reproduce the claims made in the paper. There could be many subjective reasons behind the inability to reproduce a scientific paper. The field of Machine Learning (ML) faces a reproducibility crisis, and surveying a portion of published articles has resulted in a group realization that although sharing code repositories would be appreciable, code bases are not the end all be all for determining the reproducibility of an article. Various parties involved in the publication process have come forward to address the reproducibility crisis and solutions such as badging articles as reproducible, reproducibility checklists at conferences (*NeurIPS, ICML, ICLR, etc.*), and sharing artifacts on *OpenReview* come across as promising solutions to the core problem. The breadth of literature on reproducibility focuses on measures required to avoid ir-reproducibility, and there is not much research into the effort behind reproducing these articles. In this paper, we investigate the factors that contribute to the easiness and difficulty of reproducing previously published studies and report on the foundational framework to quantify effort of reproducibility.

KEYWORDS

Effort of reproducibility, reproducibility, replicability, computational reproducibility, scholarly communication, science of science

1 INTRODUCTION

Scientific inquiry is established upon the pillars of community-driven reinforcing actions to establish confidence in published research [9]. In order to foster a community that relies on refined levels of confidence and trust in scientific inquiry, a methodology must exist for validating this scientific rigor and certainty in results. Estimating the effort of reproducibility can be considered a fundamental unit of assessing scientific inquiry amongst published research. Current literature on terminologies associated with reproducibility [2, 10, 20] fall short of capturing the full spectrum of signals capable of encapsulating the effort behind reproducing a scientific work. *Effort of reproducibility* is an important area worth exploring because it coalesces human cost, scientific validity, and confidence levels in scientific rigor into an impactful concept. Its existence, therefore, serves a greater good in scientific communities and is helpful in picking broad-spectrum signals when discussing cost, validity, and confidence associated with efforts to reproduce existing scientific work.

When researchers across the social sciences sounded distress signals under the broader cause of *reproducibility crisis [1, 8, 22]*, it set forth momentum for researchers across various computational science disciplines to self-introspect and assess the extent of this reproducibility crisis. Additionally, entities such as journals, conferences, and academic and peer-review communities started taking note of the crisis, and released protocols [4], policies [21], and checklists [11]. To an extent, these measures of caution can safeguard future research from a range of reproducibility-related issues. Moreover, having checks and balances to foster a culture of reproducible research can sustain the trust and confidence the general public places in science. However, these actions cannot do justice to capture the human effort behind the journey of reproducibility. More importantly, if there exists a rubric to encapsulate the effort of reproducibility, we could benefit from including it in our existing discussions of checklists, protocols, and policies for ensuring reproducibility.

Large-scale efforts such as the *Reproduciblity Project [5] and Open Science Collaboration [6]* were established to overcome the disincentives of reproducing already published work. Furthermore, the existence of open-access peer-reviewed journals such as *ReScience* [15] provided a platform to critically analyze and voice concerns when replicating already published scientific papers along with the added incentive of publishing these replication reports. These large-scale projects and open-access journals are necessary for modern science as an institution to produce reproducible research. Given our current understanding of the refined scientific process, the existence of various reproducibility checklists, and the presence of a vibrant open-access peer-review community, we have several useful ingredients for discovering answers about the effort of reproducibility.

Outlining these principles and unifying them under the framework of effort helps us contextualize the challenges associated with quantifying the effort itself. In this study, we lay a foundational groundwork for analyzing the effort of reproducing scientific articles from the research community's perspective. Data from the Machine Learning Reproducibility Challenge was collected and used for performing a.) an inductive qualitative analysis, b.) a quantitative analysis using Topic Models. The goal here is to systematically discover a distribution of factors responsible for encapsulating the effort of reproducibility.

2 RELATED LITERATURE

The motivation to quantify the state of reproducibility in computational science research [7] allowed critical dissection of the scientific method. The extent to which science is reproducible is a fundamental question in many studies such as [12, 13, 25]. Although these studies might interchangeably use the word *replicability* over *reproducibility* [2], there is a higher level of agreement on the merit of quantifying reproducibility. But conceptually, reproducibility as highlighted by[7], is "the ability of an independent research team to produce the same results using the same method based on the documentation made by the original research team."

Elaborate literature on reproducibility has provided members of the scientific community enough room to have a nuanced understanding of the terms and definitions associated with reproducibility. The new frontiers for reproducibility lie in doing something actionable with this knowledge. For this reason, we can observe plenty of interest [14, 17, 23] towards discovering, quantifying, and predicting research reproducibility at different levels such as *Methods reproducibility and Results reproducibility* [7]. Although these are significant endeavors, it is essential to discover the causal factors that contributed to the effort required to reproduce previously published work.

Our work incorporates the motivations of all of the above mentioned studies to discover reasons responsible for easing or complicating the effort of reproducibility. Additionally, we showcase the benefit of taking inductive qualitative analysis (human) and combining it with quantitative topic models (machine) feedback. Essentially, knowledge from this approach can be incorporated as priors for various downstream modeling tasks from human-in-the-loop machine learning techniques [24].

3 BUILDING THE DATASET

The Machine Learning Reproducibility Challenge from the years 2020 [19] and 2021 [18] provided a path for us to ask questions about the underlying effort in reproducing scientific articles. The primary goal of the ML Reproducibility Challenge is to have a community of researchers investigate the claims made in scholarly articles published at top conferences. The community selected papers and attempted to verify the claims made in the paper by reproducing computational experiments. The subsequent reports highlighting detailed information about the scope of reproducibility and what was easy and difficult for the researchers while replicating the original article were published on ReScience[1] [16].

ReScience is an open-access peer-reviewed journal with the goal of publishing researchers' endeavors to replicate computations of already published research using independently created, free, open-source software (FOSS). Demonstrably, the expected outcome from repeating experiments from previously published research is a verifiable status of reproducibility. Analyzing the additional information in these reports could support the research community in understanding the operational framework of effort necessary to reproduce published work.

Although the first three versions *ICLR'18, ICLR'19, NeurIPS'19* of the reproducibility challenge existed before 2020, we were interested in gathering data only from the recent editions of the ML Reproducibility Challenge (2020, 2021). The recent editions had reproducibility reports which followed a templated structure. The homogeneous nature of these reports was a motivating factor in focusing our analysis on recent editions. The combined articles from both editions (2020, 2021) were 87, of which 15 were removed because they belonged to adjacent disciplines of ML but not ML itself. Additionally, two more articles were removed from the final dataset because they were editorials of the reproducibility challenge. The final tally of articles selected from the reproducibility challenge is 70. We also got information about the original articles (i.e., authors and meta-level) from Google Scholar. We built scripts to automatically extract this information using python software libraries such as urllib3, Beautiful Soup, and BibtexParser. The preliminary dataset we constructed after systematic data collection consisted of the following aspects:

(1) **Meta information about the reproducibility report** such as author's name, title, DOI, year, volume, and issue.
(2) **Meta information about the original work** such as author's name, title, DOI, year, and Google Scholar citations.
(3) **Digital artifacts** such as OpenReview submission of the reproducibility study, PDF of the reproducibility report, and code repository of the replication.

Our code repository with every artifact, data, and experiments is available on Github [2].

4 INDUCTIVE QUALITATIVE ANALYSIS

We choose an inductive approach for potentially establishing a relationship between the text seen in reproducibility reports to the concept of effort. For our use case, a general inductive approach would mean repetitive, structured reading of the reproducibility reports to discover categories of reasons that made it easier or difficult to reproduce original works and reasons that served as limitations while evaluating the reproducibility of original works. The last component before initializing the analysis was having composite information of all the reproducibility reports (downloaded as PDFs) into a structured form. For this, we have utilized Allen-AI's Science Parse[3] software package to parse all of the sections of the reproducibility report.

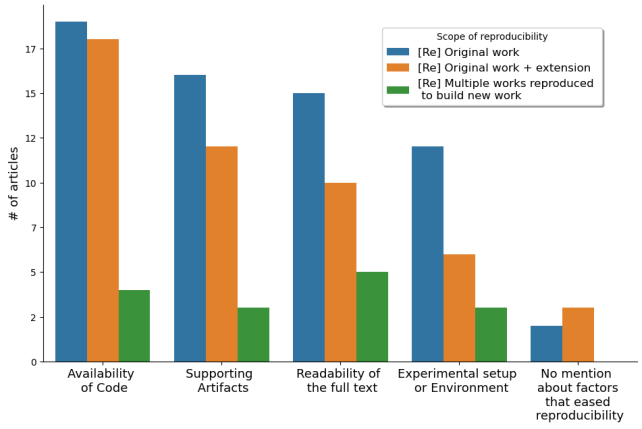

Figure 1: Reasons that eased the effort to reproduce.

4.1 Latent variables to articulate effort

Observing a general pattern amongst the reproducibility reports would require encoding sections of the report as labeled information. Almost every report from the reproducibility challenge followed a template that included sections such as the scope of reproducibility and reasons for easiness and difficulty of reproducibility. The purpose behind including these sections was to allow the community

[1] http://rescience.github.io/read/

[2] https://github.com/reproducibilityproject/effortly

[3] https://github.com/allenai/science-parse

to reflect on the individual researcher's insight while reproducing the original article. These insights included critique on elements such as clarity, thoroughness, and correctness of the original article. Therefore, it was pertinent to encode this information from the aforementioned sections as they could be latent variables acting as pointers for generalizing the effort of reproducible articles. To that extent, we extracted three sections, "Scope of Reproducibility," "What was easy," and "What was difficult" to build these latent variables.

4.2 Scope of Reproducibility

"Scope of Reproducibility" is the first section we extracted from the reproducibility report. The section's purpose is to outline main contributions of the original paper , the specific setting or problem addressed in the paper, and list the experimental methodology adopted to solve the problem. Each claim should be relatively concise; some papers may not clearly list their claims, and one must formulate them in terms of the presented experiments. The claims are roughly the scientific hypotheses evaluated in the original work. Analytically, we observed three potential categories describing the attempts to reproduce original works.

(1) **Original work**: Replications observed under this category are straightforward implementations of the original work.
(2) **Original work with supplemental extensions**: Replications of this category include implementations of the original work and additional contributions. An example of this category is the replication of the original work and the inclusion of an ablation study.
(3) **Reproducing more than one work to build new work**: Replications of this category include implementations of more than one original work to support a completely new idea.

4.3 Factors that made it easy or difficult to reproduce original study

The sections "What was easy" and "What was difficult" represent two sides of the same variable but are helpful in establishing the individual factor that is responsible for easing or burdening the effort to reproduce the research. Additionally, having two separately encoded variables (easy and difficult) highlights the existence of co-dependent factors. For instance, the availability of software artifacts might have made it easy for an individual researcher to initiate the process of reproducing a published article, but the lack of hyperparameters or data might put a strain on the effort behind reproducing the original article. Although checking for the availability of software artifacts and information about hyperparameters can be unified using the question "Is code available?", we can clearly see how the current situation cannot be boiled down into a "yes" or "no" answer.

A rigorous process of inductive encoding meant systematic reading and re-reading of the reproducibility reports to create descriptive categories that link associations to the reasons why a researcher found an original work easier or difficult to reproduce. Therefore, assumed causal reasons observed both in Fig 1 and Fig. 2 are borne out of careful human consideration.

A visual translation of our analysis can be observed in Fig 1. We can observe that the distribution of factors that made it easier to reproduce original works include *Availability of Code, Supporting Artifacts, Readability of Full text, Experimental setup or Environment*, and in some cases, *No mention about factors that eased reproducibility.*

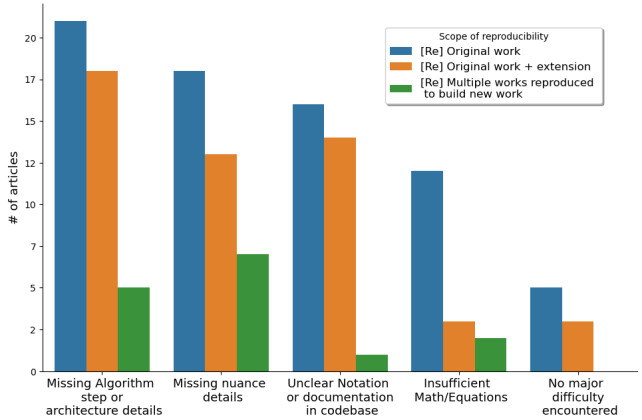

Figure 2: Reasons that made it difficult to reproduce.

Similarly, visual translation of causal reasons that made it difficult to reproduce original works can be observed in Fig 2. These factors include *Missing Algorithm step or Architecture details, Missing nuance details, Unclear notation or documentation in the codebase, Insufficient Math/Equations*, and in some cases, *No major difficulty encountered.*

Identifying reasons for easiness or difficulty is a pursuit that is both contextual and subjective. Under **contextual reasons for easiness**, Fig. 1 suggests that the *Availability of Code* and *Supporting Artifacts* are the most important contextual factors that make it easier for a researcher to replicate original works. Similarly, **contextual reasons for difficulty**, Fig. 2 show *Missing Algorithm steps or Architectural details* along with *Missing Nuance details* to be the most important factors that made it difficult for a researcher to replicate the original work.

Under **subjective reasons for easiness or difficulty**, we can notice from Fig. 1 that the *Readability of the full text* is a crucial subjective factor that eased the effort to reproduce the original study whilst under difficulty *Unclear notation or documentation in the codebase* to be a vital subjective factor that made it difficult to replicate original work. Also, it is interesting to notice a **shift in priority over reasons for easiness or difficulty** for studies whose scope is *Reproducing more than one work to build new work.* For instance, in Fig 1, the *Readability of the full text* is mentioned more as a reason than the most prevailing factor, the *Availability of Code.* Adjacently, under reasons for difficulty, as seen in Fig 2, *Missing nuance details* is the dominant factor.

4.4 Factors that limited the evaluation of reproducibility

The purpose of discovering limitations within a reproducibility study is to indicate any potential factor(s) that served as a restriction to the researcher while re-implementing the original methodology. The marginal impediments in the researcher's reproducibility journey are elemental to understanding the necessary measures taken by the researcher to overcome any limitations.

Capturing this information from ReScience reports was a starting point to understand the underlying limitations while reproducing the machine learning papers because, many a time, reproducing a machine learning study means pooling large compute resources, choosing the right hyperparameters, etc. Although these limitations are minor hurdles while replicating the original work, they do not necessarily translate into reasons for irreproducibility. Although all 70 samples in our data are successful reproductions of the original study, there still exist reasons for limitations while evaluating the original study.

Fig. 3 visualizes the distribution of factors that limited the evaluation of reproducibility, and it includes the *Necessity of Computational Resources, Missing Hyperparameters, Algorithm or Experimental Difficulty, or No mention about limitation.*

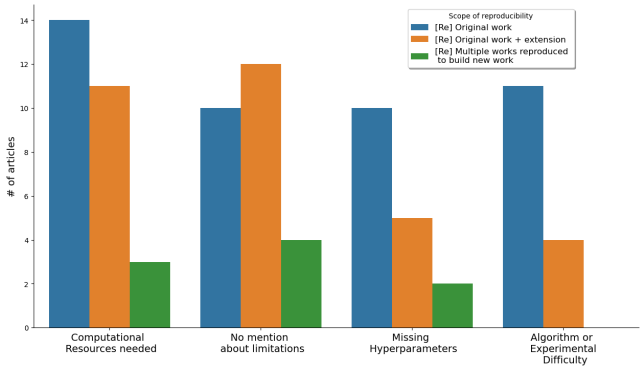

Figure 3: Reasons that served as limitations while evaluating the effort to reproduce.

The above mentioned shift in priorities for studies can be noticed in Fig 3 as well. For instance, studies under the scope **Reproducing more than one work to build new work** never mention **Algorithm or Experimental difficulty** as a subjective factor for limitation whilst studies under the scope **Original work**, consider the exact same factor to be second most important reason for limiting the evaluation of reproducibility.

5 TOPIC MODELING

Topic Models help us discover latent variables semantically related to concepts that signal the effort required to reproduce articles. Utilizing the raw texts from each report helps us preserve the distinct properties of each document whilst providing a unique representation of the mixture of reasons related to effort. Latent Dirichlet Allocation (LDA) [3] is a popular way to perform topic modeling. We utilized LDA to model reproducibility reports since those documents consist of topics. Conceptually, this is relevant to our goal, as factors that make it easier or difficult to reproduce an original work could be present as random mixtures over latent topics. We built a corpus of documents D_{easy}, and $D_{difficult}$ consisting of raw texts of "What was easy" and "What was difficult" and trained two models LDA_{easy}, and $LDA_{difficults}$. The topic coherence score for LDA_{easy} is 0.415, and $LDA_{difficult}$ 0.326. We have used the Python library **GenSim** to build the LDA models. The optimum number of topics for the LDA model was decided by ranking topic coherence scores against the number of topics.

Table 1: Most relevant terms observed by the LDA model when trained on "What was easy" corpus.

Topic	Most relevant terms
1	describe, straightforward, understand, documented
2	codebase, repository, source, instructions
3	datasets, training, scripts, experiments
4	results, ideas, evaluation, architecture
5	correspondence, addressed, peer-review, copyright

Table 2: Most relevant terms observed by the LDA model when trained on "What was difficult" corpus.

Topic	Most relevant terms
1	dataset, algorithm, implementation, method
2	training, loss, accuracy, learning
3	models, network, training, time
4	difficult, challenges, evaluation, claim
5	methods, features, performance, parameters

Table 1, and Table 2 showcase the most relevant terms observed by the LDA model when trained on their respective corpora. These terms are obtained by querying the respective topics for most dominant keywords by percentile contribution. We set a threshold of greater than 0.9 to notice the most relevant topic keywords. Analyzing the relevant terms in Table 1, we can notice close similarity towards factors mentioned in Fig 1. This outcome reinforces the relevance of our inductive encoding strategy. Interestingly, the importance of including a topic model in our study can be noticed after observing Topic five from Table 1. Thematically, it suggests **Author Correspondence** to be a dominant topic cluster observed by the LDA model as a relevant factor from "What was easy" corpus. Therefore, communication with the authors can be considered as an important factor that can ease the effort of reprocibility.

6 CONCLUSION

In this study, we lay the foundations to analyze and discover factors that can encapsulate the effort of reproducing scientific articles. Our approach of combining inductive qualitative analysis with topic modeling resulted in discovering factors that serve as reasons for easiness, reasons for difficulty, and limitations while evaluating the reproducibility of previously published work.

7 ACKNOWLEDGEMENT

This work is supported in part by NSF Grant No. 2022443.

REFERENCES

[1] Monya Baker. 2016. Reproducibility crisis. *Nature* 533, 26 (2016), 353–66.

[2] Lorena A. Barba. 2018. Terminologies for Reproducible Research. *ArXiv* abs/1802.03311 (2018).

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Ronald F Boisvert. 2016. Incentivizing reproducibility. *Commun. ACM* 59, 10 (2016), 5–5.

[5] Open Science Collaboration. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science* 7, 6 (2012), 657–660.

[6] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[7] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[8] Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

[9] National Academies of Sciences Engineering, Medicine, et al. 2019. *Reproducibility and replicability in science*. National Academies Press. https://doi.org/10.17226/25303

[10] Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (2011), 1226–1227. https://doi.org/10.1126/science.1213847 arXiv:https://www.science.org/doi/pdf/10.1126/science.1213847

[11] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *The Journal of Machine Learning Research* 22, 1 (2021), 7459–7478.

[12] Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems* 32 (2019).

[13] Edward Raff. 2021. Research Reproducibility as a Survival Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 469–478.

[14] Sarah Rajtmajer, Christopher Griffin, Jian Wu, Robert Fraleigh, Laxmaan Balaji, Anna Squicciarini, Anthony Kwasnica, David Pennock, Michael McLaughlin, Timothy Fritton, et al. 2022. A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13218–13220.

[15] Nicolas P Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A Barba, Fabien CY Benureau, C Titus Brown, Pierre De Buyl, Ozan Caglayan, Andrew P Davison, et al. 2017. Sustainable computational science: the ReScience initiative. *PeerJ Computer Science* 3 (2017), e142.

[16] Nicolas P. Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C.Y. Benureau, C. Titus Brown, Pierre de Buyl, Ozan Caglayan, Andrew P. Davison, Marc-André Delsuc, Georgios Detorakis, Alexandra K. Diem, Damien Drix, Pierre Enel, Benoît Girard, Olivia Guest, Matt G. Hall, Rafael N. Henriques, Xavier Hinaut, Kamil S. Jaron, Mehdi Khamassi, Almar Klein, Tiina Manninen, Pietro Marchesi, Daniel McGlinn, Christoph Metzner, Owen Petchey, Hans Ekkehard Plesser, Timothée Poisot, Karthik Ram, Yoav Ram, Etienne Roesch, Cyrille Rossant, Vahid Rostami, Aaron Shifman, Jemma Stachelek, Marcel Stimberg, Frank Stollmeier, Federico Vaggi, Guillaume Viejo, Julien Vitay, Anya E. Vostinar, Roman Yurchak, and Tiziano Zito. 2017. Sustainable computational science: the ReScience initiative. *PeerJ Computer Science* 3 (Dec. 2017), e142. https://doi.org/10.7717/peerj-cs.142

[17] Lamia Salsabil, Jian Wu, Muntabir Hasan Choudhury, William A Ingram, Edward A Fox, Sarah M Rajtmajer, and C Lee Giles. 2022. A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software. In *Companion Proceedings of the Web Conference 2022*. 784–788.

[18] Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Jessica Zosa Forde, Sharath Chandra Raparthy, Joelle Pineau, and Robert Stojnic. 2022. ML Reproducibility Challenge 2021. *ReScience C* 8, 2 (May 2022), #48. https://doi.org/10.5281/zenodo.6574723

[19] Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Jessica Zosa Forde, Robert Stojnic, and Joelle Pineau. 2021. ML Reproducibility Challenge 2020. *ReScience C* 7, 2 (May 2021), #1. https://doi.org/10.5281/zenodo.4833117

[20] Victoria Stodden, Friedrich Leisch, and Roger D Peng. 2014. *Implementing reproducible research*. Vol. 546. CRC Press Boca Raton, FL.

[21] Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2584–2589.

[22] Aaron Stupple, David Singerman, and Leo Anthony Celi. 2019. The reproducibility crisis in the age of digital medicine. *NPJ digital medicine* 2, 1 (2019), 2.

[23] Zhuoer Wang, Qizhang Feng, Mohinish Chatterjee, Xing Zhao, Yezi Liu, Yuening Li, Abhay Kumar Singh, Frank M Shipman, Xia Hu, and James Caverlee. 2022. RES: An Interpretable Replicability Estimation System for Research Publications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13230–13232.

[24] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* (2022).

[25] Yang Yang, Wu Youyou, and Brian Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117, 20 (2020), 10762–10768.