Domain-Specific Evaluation Strategies for AI in Journalism

Sachita Nishal*
nishal@u.northwestern.edu
Northwestern University
USA

Charlotte Li*
charlotte.li@u.northwestern.edu
Northwestern University
USA

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University
USA

1 MOTIVATION

News organizations today rely on AI tools to increase efficiency and productivity across various tasks in news production and distribution. These tools are oriented towards stakeholders such as reporters, editors, and readers. However, practitioners also express reservations around adopting AI technologies into the newsroom, due to the technical and ethical challenges involved in evaluating AI technology and its return on investments [3]. This is to some extent a result of the lack of domain-specific strategies to evaluate AI models and applications. In this paper, we consider different aspects of AI evaluation (model outputs, interaction, and ethics) that can benefit from domain-specific tailoring, and suggest examples of how journalistic considerations can lead to specialized metrics or strategies. In doing so, we lay out a potential framework to guide AI evaluation in journalism, such as seen in other disciplines [16, 38]. We also consider directions for future work, as well as how our approach might generalize to other domains.

2 EXISTING AI EVALUATION APPROACHES

2.1 From Benchmarking to Human-Centered Evaluation

Present strategies of evaluating AI models and applications range from generalized quantitative evaluation on benchmark datasets, to scoped qualitative or mixed-methods evaluation in specific contexts. In the former approach, model performance is measured on human-validated metrics, over particular benchmark datasets e.g., for image captioning [8], code synthesis [7], action detection [43], and so on. While these evaluations can be conducted rapidly and at scale, the metrics and benchmark datasets themselves capture model performance over *generalized notions of "quality" and decomposed tasks*, which limits the their capability for measuring models performance in real-world scenarios [37].

On the other end of this spectrum, human-centered and HCI-based approaches rely on *situated and contextual evaluation* of AI models and applications, such as via highly-scoped user studies (e.g. [47]), longitudinal studies (e.g. [27]), and human-grounded metrics (e.g. [18]). Recent work illustrates the importance of such scoped approaches, showing how AI models that rank lower on benchmark datasets still perform well in user studies within interactive applications [23]. While these processes allow for more nuanced evaluation in light of a particular application or context, they can be difficult to conduct in a manner that is continuous, iterative, and at scale, which would help to keep pace with model releases, and ensuing novel interactional affordances or ethical issues [13].

Frameworks for evaluating efficacy of AI models and applications within a specific *domain* (e.g. journalism, medicine, law) can help strike a balance between these approaches. For instance, domain-specific frameworks can guide crafting and validation of domain-specific task benchmarks (e.g. measuring not just "coherence" or "readability" [19] but "newsworthiness" of LLM-generated news summaries) and draw on domain-specific ethics and values for conducting ethics-based evaluations and audits (e.g. operationalizing and auditing for professional values like "immediacy" in journalism [10]). To this end, frameworks must identify domain-specific aspects, such as tasks, values, and stakeholder needs, that benchmarking must be scoped toward. They must also provide actionable guidance for continuous evaluation in real-world settings (e.g. newsrooms, hospitals). The next section highlights how this has been approached in prior work, and how journalism could benefit from such domain-specific AI evaluation frameworks.

2.2 Domain-specific Frameworks for Evaluation

Prior work on developing domain-specific evaluations of AI mainly exists in the context of healthcare [32, 38, 42] and law [16]. Similar to journalism, medical and law practitioners' concerns for the deployment of AI into real-world use cases stem from the lack of metrics for evaluating domain-specific quality and ethical alignment. In the medical domain, researchers tackle this issue by proposing a framework that can be incorporated in various stages of model development and deployment, with an emphasis on assessing ethical dimensions of models including privacy, non-maleficence, and explainability [38]. While parts of this framework can be applicable to journalism, use-cases of AI and ethical concerns in journalism differ from medicine given the public nature and the scope of potential harm caused, calling for the development of journalism-specific frameworks. Such frameworks for benchmarking can supplement qualitative evaluation methods, since they can capture certain aspects of real-world usage scenarios while allowing for iteration and suggesting potential directions for re-design.

The Partnership on AI (PAI) has offered resources to guide AI procurement and use in newsroom, which organize and categorize useful AI tools, and suggest different ways of measuring the outcomes they produce in deployment [33]. Our work builds in this direction, but offers more specific guidelines on evaluating AI for journalism for both researchers and practitioners, and with additional focus on human-AI interaction as a site of evaluation.

*Both authors contributed equally to this research.

3 BLUEPRINTS FOR AI EVALUATION IN JOURNALISM

This section presents three considerations that evaluations of AI in journalism can include and operationalize: (1) *quality of model outputs*, based on editorial goals and news values (2) *quality of interaction with AI applications*, based on needs and work processes of stakeholders (3) *ethical alignment*, based on professional values and newsroom standards. A useful framework would support practitioners in evaluating AI models and applications along these dimensions, in a manner that is flexible, iterative, and provides feedback for future designs. To actualize these domain-specific metrics, we believe methodologies that invite the collaboration between practitioners and researchers, such as co-design and participatory design, are necessary.

3.1 Quality of Model Outputs

To evaluate the quality of AI model outputs, a suite of automatic evaluation metrics and human assessments have been proposed [28]. While some of these automatic evaluation metrics can be applied to assess output quality for specific journalistic tasks such as text summarization [22, 25], machine translation [15, 34], and object detection [26], many other journalistic tasks would benefit from evaluation strategies that center more on human assessments. These human assessment evaluation methods are often generalized based on the modality of generation but are not domain-specific. For example, for generated text, human assessments tend to focus on clarity, fluency, accuracy, and coherence of the output [19]. While these criteria translate well into the domain of journalism, they overlook nuances specific to journalistic writing practice or context, such as specificity or the provision of adequate context. Similarly, for image generation, existing metrics focus mainly on image fidelity, alignment, and counting [12, 36, 41], with a few exceptions that look at social biases, robustness, and generalization [2, 9] but lack consideration of editorial judgments around e.g. image framing.

To tailor evaluation strategies towards journalistic uses of AI, researchers and practitioners can draw from news values that guide editorial decision-making. The definitions of news values remain fluid and subjective [35], but some of these can be evaluated in model outputs, for instance, in the case of AI-generated news summaries. These elements include **controversy**, **surprise**, **timeliness**, **negative or positive overtones**, and news organization's **agenda** [17]. Additionally, the ability of AI tools to support creativity is also important for reporters. Examples of creativity support includes producing **varied** but relevant outputs, and maintaining professional **tonality**.

3.2 Interactions with AI Systems

Recent work has called attention to different aspects of *in-situ interactions* with AI systems that may lend themselves well to automated evaluation metrics. These aspects include the **ease**, **enjoyment**, and **feelings of ownership** that users experience when engaged in tasks like question-answering, solving crossword puzzles, and generating metaphors with AI [23]. Human-AI interaction can also be evaluated by the *long-term goals* that a system facilitates, such as promoting **personal growth** and **emotional resilience**

for users [6]; making connections outside users' comfort zone [20]; and allowing **customization** of applications or **appropriation** for novel use-cases [1, 44].

Within journalism, the in-situ interactions and long-term processes that a system could optimize will vary by stakeholders and the tasks facilitated by AI. For instance, AI systems that provide writing feedback to reporters may be evaluated based on the **new perspectives or angles** they add to a reporter's writing [30] (e.g. by measuring semantic similarity between initial and post-feedback writing). How interactions with AI are configured also impacts the choice of metrics: minor errors in AI outputs (i.e. low **accuracy**) may be more acceptable to reporters engaged in brainstorming and ideation with text or image models [31], than to readers who only view a final product. Recent work also suggests a desire for long-term **skill development** among reporters who use AI systems [31], which could be evaluated based on periodic user surveys and AI usage analytics over time. A shared criterion of interaction quality across stakeholders could be the **enjoyment** they experience when engaged with an AI system, which applies both to reporters who use specific tools in news production (e.g. [39, 46], or to readers who rely on AI (e.g., when solving crossword puzzles [23]).

Datasets to evaluate these metrics will derive from users' interactions with AI systems, rather than solely model outputs like in Section 3.1. Understanding the short and long-term goals of different stakeholders, as well as the configurations of their interactions with AI, can support the design of such interaction metrics.

3.3 Ethical Alignment of AI Systems

Different professions and institutions adhere to different ethical principles and codes of conduct [10, 21]. Recent work suggests that subjective and multivalent **principles of journalistic practice** (e.g. truth, independence, accountability) can be translated to AI systems via value-sensitive design approaches and ethics-based audits [11]. Evaluation practices can also measure adherence to **codes of conduct and style guides** of different newsrooms to be more sensitive to institutional needs (e.g. [14, 40, 45]). This translational exercise can also provide practitioners with an opportunity to reflect on the inherent limitations of their existing codes and guides [4, 5].

Challenges to general-purpose evaluations of ethics and values hold for domain-specific evaluations as well: operationalizing and evaluating ethics can be difficult because generative AI models cannot produce consistent, causal explanations for their outputs [24]. Fine-tuned versions of the same model can also exhibit drastic variations in their adherence to ethical values. Journalism's orientation toward timeliness of communication also exacerbates these challenges. This further indicates the need for iterative and continuous evaluations of AI models and applications [29]. It also indicates a need for technical innovation to support operationalization and of journalism-specific ethics and values in AI systems.

4 FUTURE DIRECTIONS AND CONCLUSION

In this position paper, we summarized current approaches for AI evaluation, and made recommendations for additional journalism-specific evaluation criteria across three different aspects of AI evaluation. The primary goal of this paper is to call for researchers and

practitioners to come together and develop domain-specific frameworks for evaluating AI systems, so the adaptation of AI tools into newsrooms become easier and more equitable. Such frameworks can enable the development of procurement guidelines for AI tools in newsrooms, as seen in prior work [33]. We further hope this approach can empower stakeholders to create newsroom-specific custom evaluation datasets, for both short-term and longitudinal assessments of the technology.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 00 (2023), 19984–19996. https://doi.org/10.1109/iccv51070.2023.01834

[3] Charlie Beckett and Mira Yaseen. 2023. *Generating Change: A Global Survey of What News Organisations Are Doing with AI*. Technical Report. JournalismAI, London School of Economics.

[4] Lily G. Bessette, Sacha C. Hauc, Heidi Danckers, Agata Atayde, and Richard Saitz. 2022. The Associated Press Stylebook Changes and the Use of Addiction-Related Stigmatizing Terms in News Media. *Substance Abuse* 43, 1 (Dec. 2022), 127–130. https://doi.org/10.1080/08897077.2020.1748167

[5] Steve Bien-Aimé. 2016. AP Stylebook Normalizes Sports as a Male Space. *Newspaper Research Journal* 37, 1 (March 2016), 44–57. https://doi.org/10.1177/0739532916634640

[6] Alice Cai, Ian Arawjo, and Elena L. Glassman. 2024. Antagonistic AI. arXiv:2402.07350 [cs]

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs]

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. https://doi.org/10.48550/arXiv.1504.00325 arXiv:1504.00325 [cs]

[9] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3043–3054.

[10] Mark Deuze. 2005. What Is Journalism?: Professional Identity and Ideology of Journalists Reconsidered. *Journalism* 6, 4 (Nov. 2005), 442–464. https://doi.org/10.1177/1464884905056815

[11] N. Diakopoulos, C. Trattner, D. Jannach, I. Costera Meijer, and E. Motta. 2024. Leveraging Professional Ethics for Responsible AI. *Commun. ACM* 67, 2 (Jan. 2024), 19–21. https://doi.org/10.1145/3625252

[12] Tan M Dinh, Rang Nguyen, and Binh-Son Hua. 2022. TISE: Bag of metrics for text-to-image synthesis evaluation. In *European Conference on Computer Vision*. Springer, 594–609.

[13] Amitai Etzioni and Oren Etzioni. 2016. AI Assisted Ethics. *Ethics and Information Technology* 18, 2 (June 2016), 149–156. https://doi.org/10.1007/s10676-016-9400-6

[14] Paula Froke, Anna Jo Bratton, Andale Gross, Jeff McMillan, Pia Sarkar, Jerry Schwartz, and Raghuram Vadarevu (Eds.). 2022. *The Associated Press Stylebook, 2022-2024* (56th edition ed.). Basic Books, New York, NY.

[15] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics* 10 (2022), 522–538. https://doi.org/10.1162/tacl_a_00474

[16] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John J Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=WqSPQFxFRC

[17] Tony Harcup and Deirdre O'Neill. 2017. What is News? *Journalism Studies* 18, 12 (2017), 1470–1488. https://doi.org/10.1080/1461670x.2016.1150193

[18] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. https://doi.org/10.48550/arXiv.1812.04608 arXiv:1812.04608 [cs]

[19] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, 169–182.

[20] Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. 2016. Recommender Systems for Self-Actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, Boston Massachusetts USA, 11–14. https://doi.org/10.1145/2959100.2959189

[21] Bill Kovach and Tom Rosenstiel. 2021. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect* (revised and updated 4th edition ed.). Crown, New York.

[22] Kundan Krishna, Prakhar Gupta, Sanjana Ramprasad, Byron Wallace, Jeffrey Bigham, and Zachary Lipton. 2023. USB: A Unified Summarization Benchmark Across Tasks and Domains. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8826–8845. https://doi.org/10.18653/v1/2023.findings-emnlp.592

[23] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating Human-Language Model Interaction. *Transactions on Machine Learning Research* (July 2023).

[24] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941 [cs]

[25] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

[27] Tao Long, Katy Ilonka Gero, and Lydia B. Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of AI Workflows. arXiv:2402.09894 [cs]

[28] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. 2023. *The AI Index 2023 Annual Report*. Technical Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

[29] Jakob Mökander, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. 2021. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics* 27, 4 (July 2021), 44. https://doi.org/10.1007/s11948-021-00319-4

[30] Enrico Motta, Enrico Daga, Andreas L. Opdahl, and Bjornar Tessem. 2020. Analysis and Design of Computational News Angles. *IEEE Access* 8 (2020), 120613–120626. https://doi.org/10.1109/ACCESS.2020.3005513

[31] Sachita Nishal, Jasmine Sinchai, and Nicholas Diakopoulos. 2023. Understanding Practices around Computational News Discovery Tools in the Domain of Science Journalism. https://doi.org/10.48550/arXiv.2311.06864 arXiv:2311.06864 [cs]

[32] Elaine O Nsoesie. 2018. Evaluating Artificial Intelligence Applications in Clinical Settings. *JAMA Network Open* 1, 5 (2018), e182658. https://doi.org/10.1001/jamanetworkopen.2018.2658

[33] PAI Staff. 2023. *PAI Seeks Public Comment on the AI Procurement and Use Guidebook for Newsrooms*. Technical Report. Partnership on AI. https://partnershiponai.org/pai-seeks-public-comment-on-the-ai-procurement-guidebook-for-ne

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (2002),

311–318. https://doi.org/10.3115/1073083.1073135

[35] Perry Parks. 2019. Textbook News Values: Stable Concepts, Changing Choices. *Journalism & Mass Communication Quarterly* 96 (09 2019), 784–810. https://doi.org/10.1177/1077699018805212

[36] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. 2022. Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark. *arXiv* (2022). https://doi.org/10.48550/arxiv.2211.12112 arXiv:2211.12112

[37] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. https://doi.org/10.48550/arXiv.2111.15366 arXiv:2111.15366 [cs]

[38] Sandeep Reddy, Wendy Rogers, Ville-Petteri Makinen, Enrico Coiera, Pieta Brown, Markus Wenzel, Eva Weicken, Saba Ansari, Piyush Mathur, Aaron Casey, and Blair Kelly. 2021. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health & Care Informatics* 28, 1 (2021), e100444. https://doi.org/10.1136/bmjhci-2021-100444

[39] L.M. Retegui. 2021. Metrics at Work: A Case Study about the Tensions in the Media Industry. *Estudios Sobre el Mensaje Periodistico* 27, 4 (2021), 1205–1214. https://doi.org/10.5209/ESMP.71296

[40] Reuters News Agency. 2021. Standards, Values & Style Guide. *Reuters News Agency* (2021).

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[42] Ian Scott, Stacey Carter, and Enrico Coiera. 2021. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health & Care Informatics* 28, 1 (2021), e100251. https://doi.org/10.1136/bmjhci-2020-100251

[43] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. 2019. Only Time Can Tell: Discovering Temporal Data for Temporal Modeling. https://doi.org/10.48550/arXiv.1907.08340 arXiv:1907.08340 [cs]

[44] Gunnar Stevens, Volkmar Pipek, and Volker Wulf. 2010. Appropriation Infrastructure: Mediating Appropriation and Production Work. *Journal of Organizational and End User Computing* 22, 2 (April 2010), 58–81. https://doi.org/10.4018/joeuc.2010040104

[45] The New York Times. 2018. Ethical Journalism. *The New York Times* (Jan. 2018). https://www.nytimes.com/editorial-standards/ethical-journalism.html

[46] N. Thurman, K. Dörr, and J. Kunert. 2017. When Reporters Get Hands-on with Robo-Writing: Professionals Consider Automated Journalism's Capabilities and Consequences. *Digital Journalism* 5, 10 (2017), 1240–1259. https://doi.org/10.1080/21670811.2017.1289819

[47] Ruiyun Xu, Yue (Katherine) Feng, and Hailiang Chen. 2023. ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4498671

This figure "acm-jdslogo.png" is available in "png" format from:

http://arxiv.org/ps/2403.17911v1