Dimensions of Generative AI Evaluation Design

P. Alex Dow **Jennifer Wortman Vaughan** **Solon Barocas** **Chad Atalla**
Alexandra Chouldechova **Hanna Wallach**
Microsoft Research
{alex.dow, jenn, solon, chad.atalla, alexandrac, wallach}@microsoft.com

Evaluating the capabilities and risks of generative AI (GenAI) models and systems is crucial for their successful development, deployment, and adoption. Despite this, many would likely agree with *New York Times* columnist Kevin Roose's recent characterization of the current state of GenAI evaluation as "a mess—a tangle of sloppy tests, apples-to-oranges comparisons and self-serving hype" [16]. Even benchmarks, which have been viewed as the gold standard for measuring progress in GenAI model development, are coming under criticism. Writing in the context of responsible AI, the 2024 AI Index Report identified "a significant lack of standardization in responsible AI reporting" wherein leading model developers test their models against different benchmarks, which the authors suggest "may reflect existing benchmarks becoming quickly saturated, rendering them ineffective for comparison" [11]. Furthermore, as NIST has observed, "measuring risk at at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage," such as when a model becomes integrated into a system [13]. All in all, there are few principles or guidelines to ensure evaluations are effective, often making it unclear which approaches are suitable for which specific evaluation objectives. What GenAI evaluation lacks, in other words, is a systematic understanding of *evaluation design*. To help address this gap, we propose a set of general dimensions that capture critical choices involved in GenAI evaluation design. Situating evaluations within these dimensions draws attention to important choices that might otherwise be overlooked—either when designing evaluations or evaluating them.

Although several prior papers have provided useful decompositions for understanding GenAI evaluations, they are limited to specific types of evaluations, or evaluations of particular concepts. This includes work on disaggregated evaluations [1], sociotechnical safety evaluations [17], evaluations with respect to the "socio-technical gap" [10], human interaction evaluations [6], and LLM benchmarks [4]. We argue that key insights can be drawn by considering a set of general dimensions that are relevant to evaluations of *any* GenAI model or system with respect to *any* concept related to its capabilities or risks. The concept of interest, such as reasoning ability or stereotyping, is crucial, and numerous taxonomies of capabilities and risks provide different options to consider [e.g., 2, 3, 8, 9]. Just as important is the object of the evaluation—be it a model, a system, or a component thereof. However, once the concept and object have been specified, an evaluation designer is left with multiple critical choices about the most appropriate methodology for evaluating that object with respect to that concept.

Building on Barocas et al.'s work on disaggregated evaluations [1], we propose a preliminary set of general dimensions that capture critical choices involved in GenAI evaluation design and are relevant to evaluations of any GenAI model or system with respect to any concept. As shown in the left side of Table 1, these dimensions include the evaluation setting, the task type, the input source, the interaction style, the duration, the metric type, and the scoring method. We arrived at these dimensions by examining numerous evaluations, though they are not exhaustive and alternatives are likely possible.

To illustrate the utility of this set of general dimensions, consider evaluating the fairness of a GenAI system. The concept of interest in such an evaluation might relate to events that compromise fairness, such as system outputs that stereotype, demean, or erase certain social groups, or to the negative fairness-related impacts of those events, such as reinforcing unjust social hierarchies. The particular concept chosen will affect the methodology that is most appropriate for evaluating the system. The dimensions can help evaluation designers at this stage by drawing attention to various important choices that might otherwise be overlooked. If we are evaluating the system with respect to system outputs that exhibit stereotyping, the evaluation might take place in a lab setting during a single session with

NeurIPS 2024 Workshop on Evaluating Evaluations (EvalEval). (NeurIPS 2024).

Dimension	Description & Example Values	RAND	OpenAI	Google
Evaluation setting	The setting in which the evaluation will take place, e.g., computer lab, wet lab, field test, production deployment	Computer lab	Computer lab	Computer lab
Task type	The type of task involved in the evaluation, e.g., multiple choice questions, objective open-ended task, subjective open-ended task, real-world action facilitation	Subjective open-ended	Objective open-ended	Subjective open-ended
Input source	The source of inputs to the model or system, e.g., human evaluator, human user/subject/expert, real world, AI	Human subject	Human subject	Human evaluator
Interaction style	The style of interaction with the model or system, e.g., single turn, iterative	Iterative	Iterative	Single turn
Duration	The duration of the evaluation, e.g., single session, longer duration, longitudinal	Longer duration	Single sitting	Single sitting
Metric type	The type of metric, e.g., incidence, performance, feasibility, relative incidence/performance/etc.	Relative feasibility	Relative performance	Incidence
Scoring method	The method for scoring outputs or behaviors, e.g., automated, human expert	Human expert	Human expert	Human expert

Table 1: Left: Our proposed set of general dimensions that capture critical choices involved in GenAI evaluation design. Right: How three GenAI evaluations of biological threats, conducted by RAND [12], OpenAI [14], and Google DeepMind [15], might be situated within these dimensions. (Shaded cells for a dimension indicate evaluations that have the same value for that dimension.)

inputs obtained from a variety of sources. However, if we are evaluating the system with respect to reinforcing unjust hierarchies, longitudinal field tests might be more appropriate with inputs provided by humans in iterative interactions with the system. In some cases, multiple complementary evaluations that represent different points on these dimensions may be needed to shed light on a particular concept.

As a non-hypothetical example, consider three real-world GenAI evaluations of biological threats. The U.S. Department of Homeland Security (DHS) [2024] noted that GenAI systems can lower barriers to entry for chemical, biological, radiological, and nuclear (CBRN) attacks and called for "a standard framework... for pre-release evaluation and red teaming of AI models." Although no such standard framework currently exists, there is a growing body of work on evaluating GenAI models and systems with respect to CBRN threats. Table 1 shows our codings for how three recent GenAI evaluations of biological threats, conducted by RAND [12], OpenAI [14], and Google DeepMind [15], are situated within the dimensions. Several insights emerge from this: (A) All three evaluations used human expert scoring, which may indicate that human judgment is essential for assessing complex biological threats, or it could suggest the need to explore alternative scoring methods. (B) RAND and OpenAI's evaluations both involved human subjects generating inputs to a GenAI system with iterative interactions, while Google DeepMind's human evaluators worked with experts to create adversarial inputs for prompting a GenAI model with single-turn interactions. The former seems more consistent with the concerns DHS [5] voiced about lowering barriers to entry for human actors, while the latter might offer a more direct assessment of the model, avoiding the noise introduced by the variation in human subjects. (C) While all three evaluations used different metric types, both RAND and OpenAI's evaluations used relative metrics, comparing outputs produced by multiple sets of human subjects, some with only access to the internet and others that also had access to a GenAI system. This choice aligns with calls to assess the marginal risks of AI [7]. In contrast, Google DeepMind's evaluation, which is still under development, focused on the incidence of "problematic" outputs [15].

As the above examples illustrate, situating evaluations within our proposed set of general dimensions can draw attention to important choices that might otherwise be overlooked—either when designing evaluations or evaluating them. Specifically, the dimensions can guide decision-making during GenAI evaluation design, helping evaluation designers determine the most appropriate methodology for evaluating a particular GenAI model or system with respect to a particular concept. The dimensions can also provide a structure for comparing different evaluations. Given the critical role of evaluations in the successful development, deployment, and adoption of GenAI models and systems, we hope that our proposal encourages a more methodical and explicit approach to GenAI evaluation design.

Broader Impacts

By situating GenAI evaluations within a set of general dimensions, this paper seeks to address the current lack of principles or guidelines for GenAI evaluation design. We hope this will lead to a number of positive impacts, including improving evaluations, enhancing our collective understanding of the capabilities and risks of GenAI models and systems, and supporting their responsible development, deployment, and adoption. However, there are also potential negative impacts that must be considered and mitigated. First, there is a risk that the relatively small number of dimensions we proposed could give an illusion of simplicity, leading evaluation designers and other stakeholders to miss the necessary nuance involved in designing and conducting effective evaluations. Similarly, if these dimensions prove generally useful, there is a risk of overfocus, causing evaluation designers to overlook other important choices that are not captured by them. As we further expand on the work described in this paper, we can mitigate these potential negative impacts by articulating clear guidelines for their use and non-use, as well as encouraging the development of alternative structures for GenAI evaluation design.

Limitations

This paper reflects a work in progress. It is possible that particular types of GenAI evaluations require additional dimensions that we have not yet identified. In future work, we will seek to identify any gaps in our proposed set of general dimensions and address them to improve comprehensiveness.

To enhance the practicality of the dimensions, it would be beneficial to provide concrete examples of what can and cannot be learned from a GenAI evaluation given a particular choice. Such examples would help evaluation designers understand the implications of their choices and improve the overall utility of the dimensions. Additionally, to gain a deeper insight into the effectiveness of the dimensions, it is essential to apply them to a broader spectrum of evaluations across various concepts and objects. This expanded application would allow us to discern patterns and make further refinements.

Finally, although we demonstrated how to examine existing GenAI evaluations, such as three real-world GenAI evaluations of biological threats, we touched only briefly (and hypothetically) on using these dimensions to design new evaluations. This gap highlights the need for further research.

References

[1] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 368–378, 2021.

[2] Su Lin Blodgett. *Sociolinguistically driven approaches for just natural language processing*. PhD thesis, University of Massachusetts Amherst, February 2021.

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP, 2020. URL https://arxiv.org/abs/2005.14050.

[4] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[5] DHS. Department of Homeland Security report on reducing the risks at the intersection of artificial intelligence and chemical, biological, radiological, and nuclear threats, April 2024. URL https://www.dhs.gov/sites/default/files/2024-06/24_0620_cwmd-dhs-cbrn-ai-eo-report-04262024-public-release.pdf. Accessed: 2024-09-16.

[6] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static AI evaluations: Advancing human interaction evaluations for LLM harms and risks. *arXiv preprint arXiv:2405.10632*, 2024.

[7] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*, 2024.

[8] Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. Taxonomizing and measuring representational harms: A look at image tagging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14277–14285, 2023.

[9] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

[10] Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*, 2023.

[11] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The AI index 2024 annual report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2024.

[12] Christopher A. Mouton, Caleb Lucas, and Ella Guest. *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach*. RAND Corporation, Santa Monica, CA, 2023. doi: 10.7249/RRA2977-1.

[13] NIST. Artificial intelligence risk management framework (AI RMF 1.0), 2023. URL https://doi.org/10.6028/NIST.AI.100-1. Accessed: 2024-09-06.

[14] Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn (Froggi) Jackson, Steven Adler, Rocco Casagrande, and Aleksander Madry. Building an early warning system for LLM-aided biological threat creation, 2024. URL https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/. Accessed: 2024-09-16.

[15] Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating frontier models for dangerous capabilities. 2024. URL https://arxiv.org/abs/2403.13793.

[16] Kevin Roose. A.I. has a measurement problem. *The New York Times*, April 2024. URL https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html. Accessed: 2024-09-05.

[17] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*, 2023.