Ethical Considerations of Benchmarking

Victor Kariofillis, Jingyang Liu, Natalie Enright Jerger
University of Toronto
{viktor.karyofyllis, jingyang.liu}@mail.utoronto.ca, enright@ece.utoronto.ca

***Abstract*—Rapid advancement in the computer industry has sparked growing concerns about ethical aspects of computing. In this position paper, we explore an overlooked area: ethical dimensions of benchmarking practices in computer architecture. The selection of benchmarks embeds underlying ethical values into the final design. In light of this, we identify and discuss various shortcomings in current benchmark practices, point out their ethical implications, and make several proposals for how the computer architecture field can address them.**

I. Introduction

Recently, the field of computing has increasingly acknowledged and investigated the societal impact and ethical implications of computer systems. For example, sustainability – in terms of carbon emissions – is gaining traction in both academia and industry [1]–[3]. At the level of applications and algorithms, ethical considerations have also sparked a wide range of research and discussion [4], [5]. However, we identify a crucial area in the architecture community that has not yet been considered from an ethical lens. Namely, the risk that our selection of benchmarks is perpetuating biases and inequalities. As the choice of benchmarks can significantly influence design decisions, it is imperative that benchmarks suites be ethically constructed. In this work, we explore the ethical dimensions of benchmarking in computer architecture, discuss several shortcomings in our current practices, and provide proposals to address these deficits.

II. Benchmark Practices

To guide our discussion, we start by looking at a related field where ethical considerations have received much more attention, namely AI. We discuss inadequate representation and diversity, biases in datasets, and the overemphasis on performance within benchmarks. We then shift our focus to computer architecture. Additionally, we touch upon initiatives within the AI field aimed at tackling these concerns.

A. AI Benchmark Concerns and Initiatives

Current benchmarks in AI present several challenges. A significant concern is the lack of diversity in datasets and models, particularly in Natural Language Processing (NLP) tasks. Most benchmarks [6]–[11] and models [12]–[14] focus on only a few languages [15] out of the over 6500 existing today [16]. This limited scope overlooks the fundamental differences between languages. The resulting models may be fundamentally different and would need to be evaluated.

Moreover, biases are deeply ingrained in AI models and datasets, reflecting the values and perspectives of their creators [17]. These biases are often not explicitly addressed during the peer-review process or after publication. This was highlighted by a recent analysis of highly-cited machine learning publications [18]. The prevailing values in these papers prioritize generalization, efficiency, interpretability, and novelty, with minimal consideration for ethics-related values like bias elimination. This issue has gained prominence, in light of representation concerns highlighted by Google's Gemini model [19].

Another prevalent concern with current AI models lies in their predominant emphasis on performance metrics. The pursuit of higher performance often correlates with the size of models, given contemporary paradigms in deep learning research. Achieving superior performance is heavily reliant on access to extensive datasets and costly computational resources. This trajectory appears increasingly unsustainable from both economic and environmental perspectives [4], [20].

While ethical concerns persist in AI, significant strides have been made to tackle them. There have been numerous publications proposing solutions like strategies to mitigate biases, the establishment of accountability frameworks, and review processes overseen by institutional review boards (IRB) [21], [22]. Both public entities and private companies have issued documents and guidelines outlining ethical standards for AI [23]. For instance, the Conference on Neural Information Processing Systems (NeurIPS) mandates authors to complete a checklist on how their submission addresses the broader societal impacts of their research [24].

B. Computer Architecture Benchmark Concerns

Issues in computer architecture benchmarks mirror those found in the AI domain. In this section, we outline ethical and inclusivity issues, including inadequate representation, offensive content, biased workload prioritization, and growing complexity. Furthermore, we underscore the importance of conducting thorough evaluations and fostering heightened ethical awareness within the field.

Computer architecture benchmarks exhibit a lack of diverse workloads, failing to encompass a wide range of real-world applications and usage scenarios. For instance, mobile benchmark suites gather popular apps to represent common tasks [25]–[28]. However, it is unclear if these apps reflect usage by people from different cultural, age-diverse and socioeconomic backgrounds. Studies have shown that mobile app usage varies widely by country, influencing how often apps are used, which ones are popular, and even how much users spend on them [29]. For example, app users in Russia, Mexico, China, and India demonstrate a higher propensity to invest in apps, driven by the perception that paid apps offer superior quality compared to users in Canada, Australia, Germany, and the United Kingdom. Additionally, benchmarks often overlook the possibility of unreliable internet connections, which are common in many regions. They abstract the network component, replacing it with local data [26], [27]. Local data more closely mimics a reliable, high-speed internet connection which is more likely to be the case in developed economies.

Furthermore, the presence of offensive or controversial content within benchmarks raises ethical and inclusivity concerns. This is evident with the Lenna picture and its extensive usage in image processing research [30]. It has faced criticism due to its problematic origin and objectification of the model. This example has raised concerns about the appropriateness of using such images for evaluation purposes in the modern context of inclusivity and respect. The use of such images also results in the alienation of women from the field. As is noted on the Journal of Modern Optics, "(w)hatever its merits, the Lenna image's origin is incompatible with our community's sincere attempts to encourage diversity and respect in Science and Technology" [31].

Another significant challenge is the tendency of computer architecture benchmarks to prioritize workloads relevant to engineers, neglecting the diverse needs of other fields. For instance, supercomputers are often evaluated using numerical linear algebra-based benchmarks [32], [33], which may not reflect the real-world requirements of non-engineering applications. This approach inadvertently undervalues advancements in areas outside traditional engineering domains.

Moreover, benchmarks have grown substantially in size and complexity over time. For example, the SPEC CPU 2017 suite has doubled in code size compared to its 2006 counterpart, leading to longer execution times and impracticality in using the entire suite [34], [35]. This growth in size often emphasizes compute-heavy benchmarks while overlooking crucial factors like energy efficiency and sustainability. Although specialized suites like SERT [36] from SPEC focus on energy efficiency, there is a pressing need to integrate such considerations into widely used benchmarks like SPEC CPU to ensure a more comprehensive evaluation of computer architecture designs.

While efforts have been made in the AI domain to tackle ethical issues, the same level of attention has not been given to addressing these concerns in computer architecture. It is essential to start similar endeavors within computer architecture research to effectively handle biases and improve ethical considerations in benchmarking practices.

III. Proposals

In this section, we make some proposals to address the issues we highlight previously.

Proposal 1: Technical and ethical evaluations. It is crucial to go beyond traditional metrics like performance and energy efficiency and consider metrics that assess cost-effectiveness, sustainability, and accessibility. Similarly, addressing biases in benchmarks is crucial for developing a comprehensive evaluation framework. Guidelines must be established for researchers to adhere to during benchmark creation and evaluation. Implementing bias mitigation strategies, (e.g., preprocessing methods that entail identifying and rectifying biases in the data prior to model training [37]), can empower researchers to make informed decisions about benchmark suite development. For instance, when designing a benchmark with facial recognition workloads, researchers following diversity guidelines can ensure the representation of all skin tones [38].

Proposal 2: Foster diversity in benchmarks. Future benchmarks and research should prioritize creating diverse datasets across multiple dimensions. This diversity should span languages, cultures, and perspectives [39]. By including a broader range of user groups, benchmarks become more inclusive and representative. Additionally, incorporating workloads from various scientific disciplines into benchmark suites is also essential. Diverse datasets not only promote understanding of different linguistic and cultural contexts but also enhance the relevance and impact of benchmarks. Furthermore, the integration of more diverse benchmarks may lead to significant variations in resultant hardware designs.

Proposal 3: Multi-disciplinary research is necessary. Recognizing the limitations of our expertise, it is crucial to engage with diverse fields to encourage idea exchange and develop innovative approaches to benchmark development. In AI, audits involving ethicists, domain experts, and diverse stakeholders lead to independent reviews that enhance the evaluation of the ethical implications of algorithms [21]. A comparable approach should be adopted in computer architecture to create ethical frameworks that researchers can follow.

Proposal 4: Study the influence of sponsorships. While industry sponsorships are commonly acknowledged in publications, there has been limited investigation into their potential impact on benchmark selection and usage. Studying sponsorship would explore whether certain benchmarks favoured by industry receive disproportionate attention. One consequence of this trend could be the neglect of other benchmarks that might provide a broader representation of various usage scenarios. By analyzing patterns in benchmark creation and utilization, researchers can evaluate the extent to which industry interests guide research agendas in computer architecture. Such scrutiny would promote increased transparency and accountability within research practices.

IV. Conclusion

The computer architecture community faces significant ethical challenges that demand proactive and interdisciplinary solutions. We need to recognize the scope of our capabilities in resolving ethical concerns and extend our efforts across disciplinary lines to address them effectively. It is crucial to recognize that the value of our work is inherently intertwined with ethical considerations. As Ben Green notes, "broad cultural conceptions of science as neutral entrench the perspectives of dominant social groups, who are the only ones entitled to legitimate claims of neutrality" [40]. Therefore, embracing diversity, equity, and inclusion in our benchmarking practices is not only a technical imperative but also an ethical imperative that shapes the future trajectory of computer architecture research and development.

Acknowledgements

We thank the members of the NEJ group for their feedback on this work. We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2020-04179. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program and through the support of the University of Toronto McLean Award.

References

[1] L. Eeckhout, “Focal: A first-order carbon model to assess processor sustainability,” in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2024.

[2] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu, “Chasing carbon: The elusive environmental footprint of computing,” in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 854–867.

[3] B. Acun, B. Lee, F. Kazhamiaka, K. Maeng, U. Gupta, M. Chakkaravarthy, D. Brooks, and C.-J. Wu, “Carbon explorer: A holistic framework for designing carbon aware datacenters,” in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2023, pp. 118–132.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623. [Online]. Available: https://doi.org/10.1145/3442188.3445922

[5] B. Laufer, S. Jain, A. F. Cooper, J. Kleinberg, and H. Heidari, “Four years of FAccT: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 401–426.

[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” *CoRR*, vol. abs/1804.07461, 2018. [Online]. Available: http://arxiv.org/abs/1804.07461

[7] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems,” *CoRR*, vol. abs/1905.00537, 2019. [Online]. Available: http://arxiv.org/abs/1905.00537

[8] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[9] E. Filatova, “Irony and sarcasm: Corpus generation and analysis using crowdsourcing,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 392–398. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/661_Paper.pdf

[10] S. Merity, C. Xiong, J. Bradbury, and R. Socher, “Pointer Sentinel Mixture Models,” *CoRR*, vol. abs/1609.07843, 2016. [Online]. Available: http://arxiv.org/abs/1609.07843

[11] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, “Findings of the 2014 workshop on statistical machine translation,” in *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 12–58. [Online]. Available: https://aclanthology.org/W14-3302

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[13] OpenAI *et al.*, “GPT-4 Technical Report,” 2024.

[14] Gemini Team *et al.*, “Gemini: A Family of Highly Capable Multimodal Models,” 2023.

[15] D. Blasi, A. Anastasopoulos, and G. Neubig, “Systematic inequalities in language technology performance across the world’s languages,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5486–5505. [Online]. Available: https://aclanthology.org/2022.acl-long.376

[16] H. Hammarström, ““Ethnologue” 16/17/18th editions: A comprehensive review,” *Language*, vol. 91, no. 3, pp. 723–737, 2015. [Online]. Available: http://www.jstor.org/stable/24672170

[17] T. LaCroix and A. S. Luccioni, “Metaethical Perspectives on ’Benchmarking’ AI Ethics,” *arXiv preprint arXiv:2204.05151*, 2022.

[18] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, “The Values Encoded in Machine Learning Research,” *arXiv preprint arXiv:2106.15590*, 2022.

[19] T. Warren. (2024) Google pauses Gemini’s ability to generate AI images of people after diversity errors. [Online]. Available: https://www.theverge.com/2024/2/22/24079876/google-gemini-ai-photos-people-pause

[20] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “The Computational Limits of Deep Learning,” *arXiv preprint arXiv:2007.05558*, 2022.

[21] M. M. Islam and J. Shuford, “A Survey of Ethical Considerations in AI: Navigating the Landscape of Bias and Fairness,” vol. 1, pp. 1–5, 02 2024.

[22] C. E. A. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, and A. Dafoe, “Institutionalizing ethics in AI through broader impact requirements,” *Nature Machine Intelligence*, vol. 3, no. 2, p. 104–110, Feb. 2021. [Online]. Available: http://dx.doi.org/10.1038/s42256-021-00298-y

[23] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices,” 2019.

[24] “NeurIPS Paper Checklist Guidelines,” https://neurips.cc/public/guides/PaperChecklist, accessed: 2024-03-20.

[25] A. Gutierrez, R. G. Dreslinski, T. F. Wenisch, T. Mudge, A. Saidi, C. Emmons, and N. Paver, “Full-system analysis and characterization of interactive smartphone applications,” in *2011 IEEE International Symposium on Workload Characterization (IISWC)*, 2011, pp. 81–90.

[26] D. Pandiyan, S.-Y. Lee, and C.-J. Wu, “Performance, energy characterizations and architectural implications of an emerging mobile platform benchmark suite - MobileBench,” in *IEEE IISWC*, 2013.

[27] Y. Huang, Z. Zha, M. Chen, and L. Zhang, “Moby: A mobile benchmark suite for architectural simulators,” in *2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2014, pp. 45–54.

[28] S. Fan and B. C. Lee, “Evaluating asymmetric multiprocessing for mobile applications,” in *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2016, pp. 235–244.

[29] S. L. Lim, P. J. Bentley, N. Kanakam, F. Ishikawa, and S. Honiden, “Investigating Country Differences in Mobile App User Behavior and Challenges for Software Engineering,” *IEEE Transactions on Software Engineering*, vol. 41, no. 1, pp. 40–64, 2015.

[30] Wikipedia contributors, “Lenna — Wikipedia, the free encyclopedia,” 2023, [Online; accessed 11-March-2024]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Lenna&oldid=1183832167

[31] “On alternatives to Lenna,” *Journal of Modern Optics*, vol. 64, no. 12, pp. 1119–1120, 2017. [Online]. Available: https://doi.org/10.1080/09500340.2016.1270881

[32] J. J. Dongarra, P. Luszczek, and A. Petitet, “The linpack benchmark: past, present and future,” *Concurrency and Computation: Practice and Experience*, vol. 15, no. 9, pp. 803–820, 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.728

[33] “Top 500 - the list.” [Online; accessed 18-March-2024]. [Online]. Available: TOP500.org

[34] A. Limaye and T. Adegbija, “A workload characterization of the spec cpu2017 benchmark suite,” in *2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2018, pp. 149–158.

[35] R. Panda, S. Song, J. Dean, and L. K. John, “Wait of a Decade: Did SPEC CPU 2017 Broaden the Performance Horizon?” in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018, pp. 271–282.

[36] K.-D. Lange and M. G. Tricker, “The design and development of the server efficiency rating tool (SERT),” in *Proceedings*

of the 2nd ACM/SPEC International Conference on Performance Engineering, ser. ICPE '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 145–150. [Online]. Available: https://doi.org/10.1145/1958746.1958769

[37] E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies," *Sci*, vol. 6, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2413-4155/6/1/3

[38] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[39] P. P. Ray, "Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 3, p. 100136, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772485923000534

[40] B. Green, ""Good" isn't good enough," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209379533