Testing and Evaluation of Data Analytic Approaches for Nonproliferation

Proceedings of the INMM & ESARDA Joint Annual Meeting
May 23–26, 2023

Dylan Anderson
Sandia National Laboratories[*]
Albuquerque, NM, USA
dzander@sandia.gov

Scott L. Stewart
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Alexei Skurikhin
Los Alamos National Laboratory
Los Alamos, NM, USA

Karl Pazdernik
Pacific Northwest National Laboratory
Richland, WA, USA

Joel Brogan
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Nathan Martindale
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Abstract

Advanced data analytics continue to promise faster machine-assisted analyst workflows, integration of vast troves of open-source and multimodal information, and discovery of new, previously undetected events of interest in international safeguards, arms control, and nonproliferation applications. Realizing these promises requires robust analytic validation and performance measurement, which is classically reliant on a validation dataset drawn randomly from available data and withheld during development for final assessment before deployment. However, these applications are characterized by high consequences, sparse data and rare events, and significant, often nonrandom, uncertainty. Under these considerations, a classic analytic validation often lacks real-world data complexities or relies on metrics that do not reflect how analytics may be used in support of monitoring or decision-making in practice. Furthermore, the inadvertent leakage of validation information into analytic development can inflate assessed performance characteristics, yielding overly optimistic expectations for success. Finally, analytics must be continually monitored during use to ensure they continue to meet expected performance requirements, as the data being analyzed may no longer be well represented by the data used for model training. This session highlighted several case studies detailing validation failures, lessons learned, and best practices for the development of advanced data analytics for safeguards, arms control, and nonproliferation applications as well as approaches to continually monitor analytics during production.

Keywords data analytics · machine learning · data science · testing and evaluation · nonproliferation

1 Introduction

Advanced data analytics and associated concepts in big data, machine learning, and artificial intelligence (AI) are poised to greatly enhance international safeguards, arms control, and nonproliferation applications. These approaches have been successfully and regularly applied to exponentially growing open-source and proprietary multimodal information for security and commercial purposes and have been strongly recommended for adoption into nuclear monitoring and

[*] Notice: This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.

verification (1). Furthermore, advanced data analytics also promise to boost the productivity of inspectors and analysts by reducing repetitive tasks and quickly highlighting salient or anomalous information for analyst triage and review (2).

The development of advanced data analytics classically relies upon validation procedures in which evaluation data are drawn randomly from the pool of available data and withheld during analytic development (3). The evaluation data are used to assess analytic generalizability to new, unseen data and build expectations for model performance. There are variations of this approach, such as cross-validation, which is a sampling strategy to mitigate the pitfalls of splitting an already small data pool. However, this general framework and its variations do not address many of the characteristics endemic in the application spaces of interest in nonproliferation, which include sparse data and rare events of interest, nonrandom biases and uncertainty, and inordinately high consequence decisions (4; 1).

Sparsity, imbalanced class proportions, and highly specialized or nuanced subpopulation statistics make adequate sampling to form a representative evaluation set difficult or impossible. Furthermore, common metrics used to drive model selection or evaluation do not fully represent the considerations and consequences of decision-making, particularly when competing factors, such as interpretability and performance, must be considered simultaneously. Contending with these deficiencies often leads to inadvertent leakage of evaluation information into the analytic development process through continued re-evaluation, which can inflate expected performance characteristics of analytics. Moreover, these considerations continually reemerge under the auspices of monitoring deployed analytics, wherein data and concepts drift over time and space and require constant reevaluation (5; 6).

This paper describes a special session that presented three different technical approaches, highlighted briefly in Section 2, toward tackling some of these challenges. The researchers behind these approaches then convened an interactive panel session with the audience; the themes from this panel session are described in Section 3. Finally, Section 4 concludes this paper.

2 Select Technical Approaches

Three panelists were asked to present on their recent work related to the testing and evaluation of data analytic approaches for nonproliferation as part of the special session. Each panelist focused on a different area of testing and evaluation as well as different types of input modalities into deep network models. The works presented considered an overall framework that could be used for evaluating models, calibrating textual models to have better uncertainty quantification, and developing models with built-in explainability and robustness. Each work is very briefly summarized in this section.

2.1 Building Performance Evaluation Framework of Foundation Models for Nonproliferation Applications

Recent progress in AI has culminated in foundation models (FMs) that can facilitate the development of innovative approaches for nuclear verification and geographic profiling of activities of interest. FMs are large-scale deep learning neural network models (e.g., transformer models) that are trained on very large general datasets and can then be tuned to a wide range of downstream tasks with relatively little additional task-specific training. FMs have already demonstrated a huge impact in natural language processing and are increasingly used for computer vision tasks, such as image-to-text mapping, image retrieval, and tagging. Although FMs are powerful models, adapting them for the nuclear nonproliferation domain is limited by inadequate quality and variety of data, as well as a possibility of bias in the data used to train the original FM. Testing and evaluating (T&E) FMs, including uncertainty quantification, is crucial for nonproliferation applications, where challenges include the unavailability of all the modalities all the time, unequal distribution of information across modalities, and unequal distribution of annotated data across different modalities. Los Alamos National Laboratory has developed a framework for T&E of FMs against downstream tasks such as land use, scene and image classification, object detection, localization, and segmentation. These tasks are essential for the characterization of objects and activities of interest. This framework has been demonstrated for comparing a transformer model to convolutional neural networks for scene classification using satellite imagery collected over different geographic regions using T&E metrics (see Figure 1). The work is further detailed in a proceedings article (7).

2.2 Well-Calibrated Uncertainty Quantification for Language Models in the Nuclear Domain

A key component of global and national security in the nuclear weapons age is the proliferation of nuclear weapons technology and development. One important aspect of enforcing nonproliferation policy is developing an awareness of the scientific research being pursued by other nations and organizations. To support nonproliferation goals and contribute to nuclear science research, Pacific Northwest National Laboratory trained a RoBERTa deep neural language model on a large set of research article abstracts archived by the U.S. Department of Energy's Office of Science and Technical Information, and then fine-tuned this model for classification of scientific abstracts into 60 disciplines, called

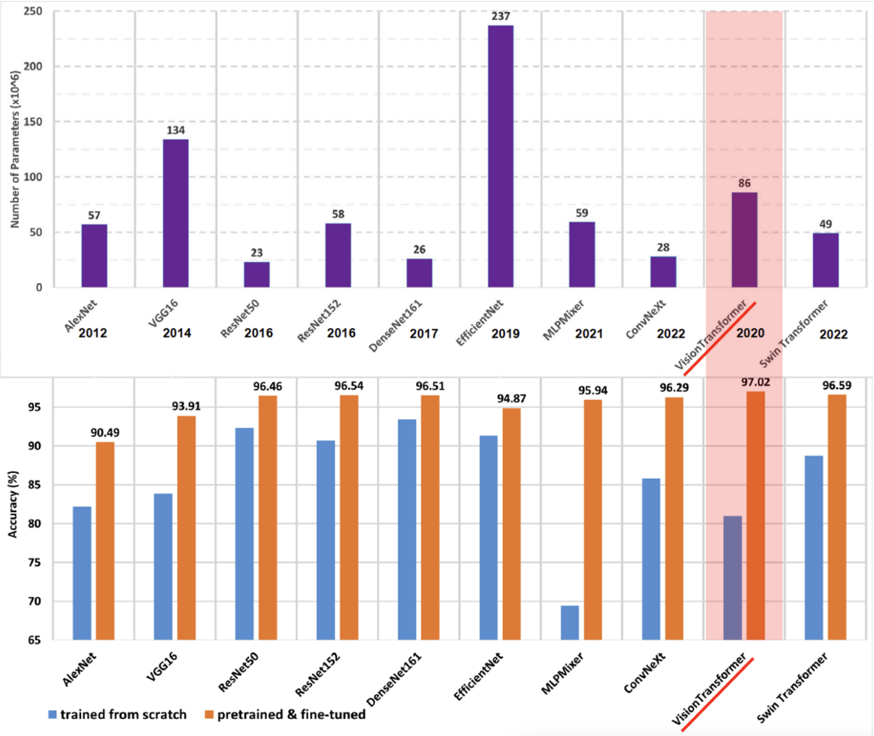

Figure 1: Testing and evaluation framework for a wide range of model scales on various downstream prediction tasks from satellite imagery.

NukeLM (8; 9). This multistep approach to training improved classification accuracy over its untrained or partially out-of-domain competitors. Classifiers must be accurate, but there is also growing interest in ensuring that classifiers are well-calibrated with uncertainty quantification that is understandable to human decision-makers. For example, in the multiclass problem, classes with a similar predicted probability should be semantically related. Therefore, they also introduce extensions of the Bayesian belief matching framework called correlated belief matching and hierarchical correlated belief matching (10). These extensions have been demonstrated to easily scale to large natural language processing models, such as NukeLM, and achieve better desired uncertainty quantification properties such as producing improved coverage rate of credible set matches (see Figure 2) and providing greater semantic interpretability by allowing for predictions over semantically similar classes.

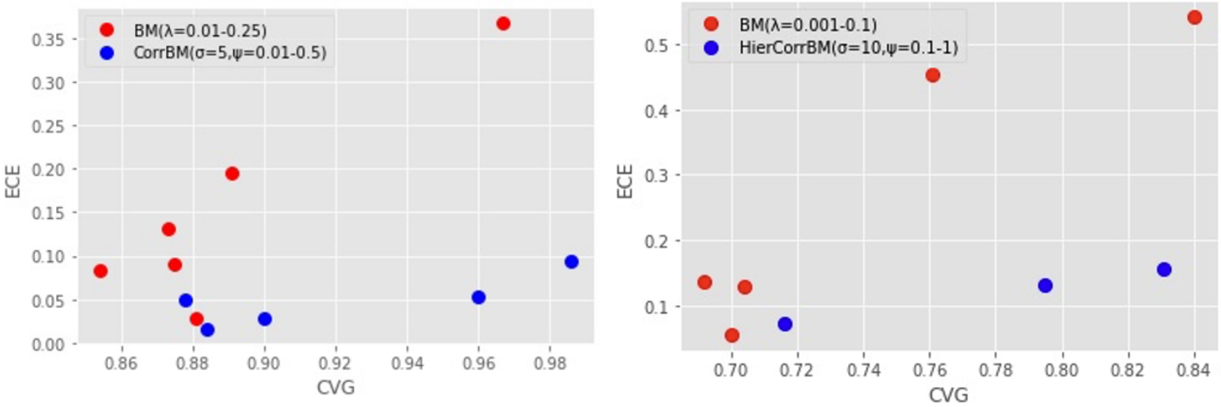

Figure 2: Correlated belief matching (left) and hierarchical correlated belief matching (right) produce better coverage than belief matching for the same expected calibration error (10).

Table 1: Neural SDEs can enhance the robustness of a base model to noise.

Noise	ConvNeXt-Base (%)	Neural SDE (%)
0.05	84.38	87.50
0.01	71.88	75.00
0.15	56.25	62.50
0.20	56.25	53.12
0.25	25.00	40.62
0.50	12.50	21.88
0.75	3.12	9.38
1.00	6.25	6.25

2.3 Confidently Explaining Spectrogram Classification Using Neural Stochastic Differential Equations

Explainable AI algorithms (XAI) can increase the confidence that a model or data analytic applied to a high-consequence application such as nonproliferation will make predictions that align to domain expertise expectations. XAI can be used to assess that a prediction was made similarly to how a domain expert would have made that same prediction. Activation maps generated from XAI are meant to indicate locations that are highly influential to the predictive output of a machine learning algorithm. These locations are called "attributes." The more an attribute fluctuates in the instance of noisy signal, the less presumed confidence in the attribute. Neural stochastic differential equations (neural SDEs) inject noise into the feed-forward prediction networks trained on spectrograms, and, along with the attribute-based-confidence approach (ABC), determine the portions of an input signal that most confidently describe what makes it unique when compared to other signals of different classes. Oak Ridge National Laboratory has shown that this neural SDE explainability approach can greatly enhance the robustness of an already high-performing base model to both random and adversarial noise (see Table 1). These explanations bring built-in robustness to model predictions and greatly increase the confidence in models deployed in new settings by ensuring they are making predictions based on evidence aligned with subject matter expert expectations.

3 Panel Discussion Themes

After the panelist presentations, the session co-chairs and audience members engaged the panelists in a conversation about testing and evaluation of approaches in this mission area. Three major themes from that conversation are presented below. The intent of this portion of the paper is not to present a solution to these issues but instead to document the necessity of consideration for these topics in future studies and endeavors.

3.1 Multimetric Optimization and Tuning

Competing demands are increasingly imposed on advanced data analytics. For example, many nonproliferation applications simultaneously require explainability and trust, uncertainty quantification, and high performance. With the emergence of FMs, demands may require high performance across a range of downstream tasks, such as predictive accuracy over different subsets or populations, or may be based upon emergent capabilities that are not well known or established during model training (11). Strategies exist for balancing optimization for multiple objectives simultaneously, but these approaches require *a priori* knowledge of relative weights or importances of objectives, or that the objectives can be combined in a manner amenable to optimization, such as a Pareto front (12). There may not be any single set of relative importance weights when an analytic is to be used by a diverse set of analysts and stakeholders.

The panelists described some possible recommendations for contending with this issue, such as setting secondary objectives to a fixed operating value and then tuning for the final primary performance objective, as in the work described in Section 2.2. In this work, techniques were tuned to produce similar expected calibration errors and then subsequently compared for uncertainty coverage and predictive accuracy (see Table 2). There are emergent benchmarks for uncertainty, such as the case study described in Section 2.1, which compared a range of models simultaneously against both accuracy and expected calibration error for the RESISC45 Satellite imagery classification dataset (see Figure 3). These nascent benchmarks allow for standardization and cross comparison of techniques against a range of metrics. However, comprehensive and computationally tractable approaches for contending with this issue, both from the algorithms and optimization as well as the metrics and benchmarks perspectives, remain elusive.

Table 2: Techniques are tuned to comparable expected calibration error operating points and then compared on other metrics. Table from work in Section 2.2.

	Belief Matching	Hierarchical Correlated Belief Matching	Dropout
Accuracy (%)	68.3	67.3	68.7
Expected calibration error	0.129	0.132	0.131
Coverage (%)	70.4	79.5	83.3
Ranking Accuracy	0.120	0.479	0.266
Sampling time (min)	2.47	1.47	216.75

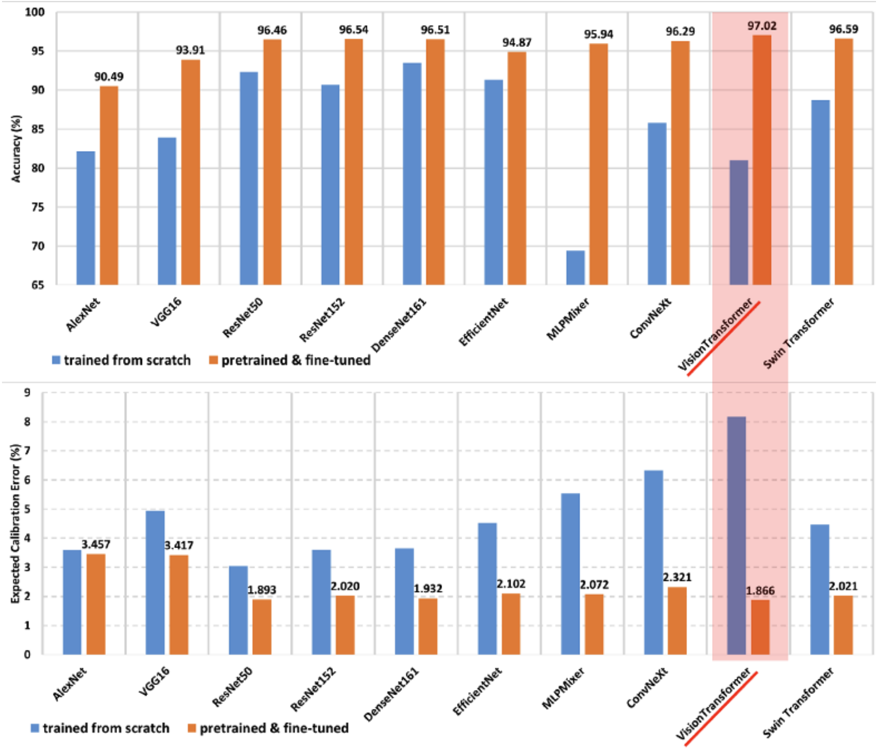

Figure 3: A new benchmark for assessing uncertainty quantification across training methods and models.

3.2 Communicating with System Users

Although many facets of interest exist for analytic performance and characterization, it remains unclear what information should be shown to users of analytics, particularly in the context of high-consequence nonproliferation applications. For example, a posterior predictive distribution remains a desirable comprehensive output for uncertainty quantification. However, it is unclear that this type of information can be used and acted upon by system users, particularly with high processing volumes. A key question was raised during the panel session: Is this an education issue for decision-makers and system users, who are increasingly immersed in machine learning technologies, or is this an algorithms development and human factors issue for system builders? The answer likely falls somewhere between and is a rapidly evolving area of importance for application of data analytics to nonproliferation.

At the same time, expectations and standards should be elucidated from subject matter experts and leveraged. In the work in Section 2.3, explanations of analytic predictions were validated with subject matter expert expectations for how predictions should be performed. Within the time-frequency domain, spectrograms should be particularly informative of specific frequency features, which would manifest in prediction explanations as predominately vertical striping along the frequency axis of input spectrograms. These expectations were used to validate the robustness of analytics and assess whether models had learned the expected features during fine-tuning. Figure 4 illustrates the difference between a model with poor attribute coherence and a reasonable attribute coherence that aligns with subject matter expert expectations.

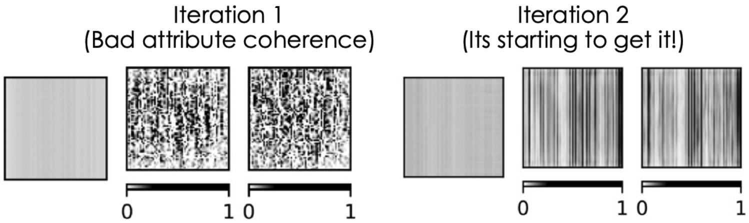

Figure 4: Prediction explanations assess the alignment between model emphasis and subject matter expert expectations.

3.3 Postdeployment Monitoring

The panel also discussed the nascent field of postdeployment and *in situ* model monitoring. That is, monitoring the health or quality of a data analytic after it has been deployed to production (6). Aspects of postdeployment monitoring are similar to the testing and evaluation techniques discussed here. For example, monitoring the distribution of predicted probabilities and continually reassessing model calibration may reveal statistical data changes that necessitate retraining or recalibration. Analytics may be deployed or available *in situ* behind application programming interfaces, which may not provide direct access to the analytic itself. In such circumstances, extrinsic assessment techniques, such as those described in the work in Section 2.1, may be required for monitoring.

Validating prediction explanations against expectations, as described in the previous section, can also help to assert that new data remains in distribution for a model's training. However, care must be taken to interpret or explain a model prediction that may be fundamentally beyond what a human can directly understand. This was evidenced in the work in Section 2.3, wherein the pretrained base ImageNet model occasionally produced saliency maps matching the expected vertical striping pattern for spectrogram classification. Careful analysis revealed that these explanations corresponded to predictions of the class "curtains" in the ImageNet dataset used for pretraining (Figure 5).

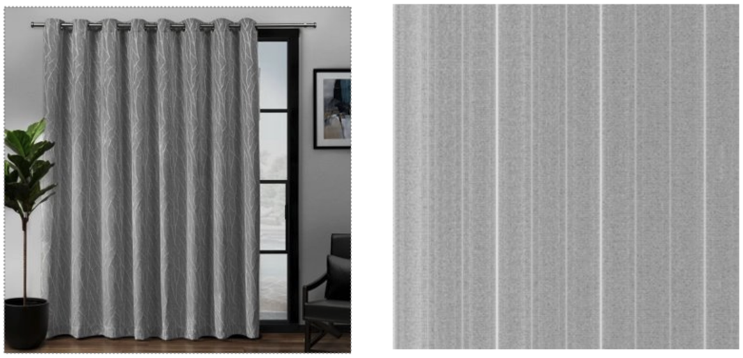

Figure 5: Explanations for predictions of "curtains" matched subject matter expert expectations for feature importance from spectrograms. Care must be taken when interpreting explanations that lie on the edge of—or beyond—human conceptualization of a problem.

4 Conclusion

This special session highlighted three different selected technical approaches for testing and evaluating data analytics for international safeguards, arms control, and nonproliferation applications. These domains possess unique characteristics, including data sparsity, rare events of interest, and exceedingly high consequences. The session also convened a panel discussion on testing and evaluation, through which themes on multiobjective optimization, communicating analytic outputs with system users, and postdeployment monitoring emerged. Panelists provided some recommendations for dealing with the associated challenges, and identified significant opportunities for impactful future research along these major themes.

5 Acknowledgments

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

This work was funded by the U.S. Department of Energy National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research and Development.

References

[1] F. J. Alexander, T. Borders, A. Sheffield, and M. Wonders, "Workshop report for next-gen AI for proliferation detection: Accelerating the development and use of explainability methods to design AI systems suitable for nonproliferation mission applications," 9 2020. [Online]. Available: https://www.osti.gov/biblio/1768761

[2] A. Vlasov and M. Barbarino, "Seven ways AI will change nuclear science and technology," *International Atomic Energy Agency*. [Online]. Available: https://www.iaea.org/newscenter/news/seven-ways-ai-will-change-nuclear-science-and-technology

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. NY: Springer New York, 8 2006.

[4] S. Stewart, D. Anderson, M. Adams, J. Dermigny, N. Martindale, K. Gabert, B. Alexandrov, L. Prasad, J. Brogan, Z. Brown, P. Bingham, and T. Grimes, "Cutting edge approaches in data analytics for nonproliferation," in *Proceedings of the 63rd Annual Meeting of the Institute of Nuclear Materials Management*, 2022.

[5] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2019.

[6] K. Pykes, "A guide to monitoring machine learning models in production," *NVIDIA Developer Blog*. [Online]. Available: https://developer.nvidia.com/blog/a-guide-to-monitoring-machine-learning-models-in-production/

[7] A. Skurikhin, G. Flynn, M. Geyer, G. Gopalan, N. Klein, J. Moore, M. Myshatyn, N. Parikh, R. Rael, S. Wanna, and E. Casleton, "Building performance evaluation framework of foundation models for nonproliferation applications," in *Proceedings of the 2nd Joint Annual Meeting of the Institute of Nuclear Materials Management and the European Safeguards Research and Development Association*, 2023.

[8] "U.S. Department of Energy Office of Scientific and Technical Information," https://www.osti.gov/, accessed: 2023-07-07.

[9] L. Burke, K. Pazdernik, D. Fortin, B. Wilson, R. Goychayev, and J. Mattingly, "NukeLM: Pre-trained and fine-tuned language models for the nuclear and energy domains," *arXiv preprint arXiv:2105.12192*, 2021.

[10] T. Joo, U. Chung, and M.-G. Seo, "Being Bayesian about categorical probability," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 July 2020, pp. 4950–4961. [Online]. Available: https://proceedings.mlr.press/v119/joo20a.html

[11] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, and et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021. [Online]. Available: https://arxiv.org/abs/2108.07258

[12] N. Gunantara, "A review of multi-objective optimization: Methods and its applications," *Cogent Engineering*, vol. 5, no. 1, p. 1502242, 2018. [Online]. Available: https://doi.org/10.1080/23311916.2018.1502242