Model-Based Data-Centric AI: Bridging the Divide Between Academic Ideals and Industrial Pragmatism

Chanjun Park[†*] **Minsoo Khang**[†], **Dahyun Kim**[†]
Upstage
{chanjun.park, mkhang, kdahyun}@upstage.ai

Abstract

This paper delves into the contrasting roles of data within academic and industrial spheres, highlighting the divergence between Data-Centric AI and Model-Agnostic AI approaches. We argue that while Data-Centric AI focuses on the primacy of high-quality data for model performance, Model-Agnostic AI prioritizes algorithmic flexibility, often at the expense of data quality considerations. This distinction reveals that academic standards for data quality frequently do not meet the rigorous demands of industrial applications, leading to potential pitfalls in deploying academic models in real-world settings. Through a comprehensive analysis, we address these disparities, presenting both the challenges they pose and strategies for bridging the gap. Furthermore, we propose a novel paradigm: Model-Based Data-Centric AI, which aims to reconcile these differences by integrating model considerations into data optimization processes. This approach underscores the necessity for evolving data requirements that are sensitive to the nuances of both academic research and industrial deployment. By exploring these discrepancies, we aim to foster a more nuanced understanding of data's role in AI development and encourage a convergence of academic and industrial standards to enhance AI's real-world applicability.

Keywords: Data-Centric AI, Model-Agnostic AI, Model-Based Data-Centric AI

1 Introduction

The divergence in the conceptualization of 'good data' between academic research and industrial application is not just prevalent but also significant, especially when considering the development and deployment of models across different domains. Academia tends to prioritize datasets that encapsulate the essence of a given task, aiming for a broad applicability that facilitates fair comparisons across different models. This approach, while valuable for theoretical exploration, often overlooks the specificities that datasets require to optimize models for specialized tasks.

In contrast, industry focuses on curating data that directly improves model performance for particular applications, integrating complex, real-world nuances including edge cases and long-tail distributions (Cha et al., 2023; Choi and Park, 2023). This discrepancy highlights that benchmark datasets, while foundational, may not suffice for the nuanced needs of practical applications (Srivastava et al., 2022), necessitating model-informed data curation to truly enhance performance (Gupta et al., 2021).

. * Corresponding Authors
† Equal Contribution

©2024 Chanjun Park and Minsoo Khang and Dahyun Kim.

Our critique extends beyond the current scope of academic data curation, which often fails to account for the dynamic interaction between data and models. This oversight is not due to a lack of research but rather to the dominant focus on data-centric AI that neglects the intricate ways in which models engage with and learn from data (Zha et al., 2023).

From an academic lens, creating datasets without considering specific model requirements might seem logical. However, for models in the industrial sphere, this practice is counterproductive. Datasets predominantly composed of straightforward positive examples do little to challenge or improve models, leading to marginal gains after reaching a certain threshold of data quantity (Soviany et al., 2022). The complexity and depth of models necessitate datasets that not only cover a wide range of scenarios but also include challenging, model-specific instances to drive significant performance improvements.

Therefore, we propose a pivotal shift away from *model-agnostic* data curation towards a *Model-Based Data-Centric AI*. This paradigm shift is essential for aligning academic research with industrial objectives, ensuring that datasets are not only theoretically comprehensive but also practically effective in enhancing model performance. By integrating model-specific insights into the data curation process, we can bridge the existing gap, enabling models to achieve their full potential in both research and real-world applications.

Moreover, it is important to note that these considerations are equally applicable to the realm of Large Language Models (LLMs) (Zhao et al., 2023). As evidenced by works such as (Madaan et al., 2023), the nuanced data requirements and model-specific considerations discussed herein are not confined to traditional AI models but extend to LLMs as well. Given their vast capabilities and varied applications, LLMs stand to benefit significantly from data curation practices that are informed by both the complexity of real-world data and the specific demands of the tasks they are designed to perform. This inclusive approach highlights the universal importance of sophisticated data curation strategies, underscoring the need for continuous evolution in our approach to data, to meet the demands of the ever-advancing AI landscape, including the burgeoning field of LLMs.

2 Data-Centric AI and Model-Specific

In the field of academia, datasets are often curated in a model-agnostic manner, to facilitate fair evaluation over various models. This paper argues that model-dependent features need to be considered in the curation process as well to produce the best performing *industrial* models instead. Two main features to be discussed in this paper are: approach-specific information and model guided data synthesis.

Approach-specific information. In curating datasets, annotating additional information that is not inherently fundamental to the task at hand can be beneficial for certain approaches. These additional annotations can range from metadata to additional labels depending of which can all benefit the downstream task solution despite not being essential to the target task.

For instance, let us discuss *Document Visual Question Answering* (DocVQA) task (where document images, text transcriptions, and question-and-answer pairs are given) as an example (Mathew et al., 2021), where two broad groups of solutions exist: *extractive* and *generative*. Generative approaches aim to generate the answer for the given question conditioned on the input image. On the other hand, extractive methods assume that the

answer to the given question is a substring of the text transcription and aim to extract or *index* the answer from the given text transcription conditioned on the input image. Despite the high output capacity of generative approaches, extractive approaches are often preferred in use-cases where the tolerance for *hallucinated* answers is low.

For generative approaches, document images, text transcriptions, and question-and-answer pairs are sufficient in providing the model with the necessary supervisory signals. Extractive approaches, on the other hand, require both the start and end indices of answer substrings in the given text transcriptions. Without these indices, extractive models cannot receive the supervisory signals for training, and the absence of such approach-specific information results in the need to re-label the dataset, that is not only costly but potentially impractical. Provisioning of such approach-specific information, beyond just task-specific ones, need to be thoroughly considered during the dataset curation stage, especially in the context of *industrial* applications.

Model guided data synthesis. Synthetic data generation is often used alongside real data curation to supplement the data deficiency and class imbalance issues (Johnson and Khoshgoftaar, 2019; Narayanan et al., 2022). While these techniques behind data synthesis are helpful in general, it is often difficult to *evaluate* the quality of the synthetic data aside from hand-picked qualitative comparisons. We argue that the impact of the synthetic data on the downstream model performance should be used as guidance in gaining insight on the quality of the synthetic data and potential improvements.

Taking character recognition task (where word images and corresponding word texts are given) as an example, character recognition in agglutinative languages such as Korean is often challenged with handling characters with long-tailed distribution (Park et al., 2023). A simple data synthesis method would be to synthesize more words using the tail part of the character distribution, *i.e.,* up-sampling the tail characters. While such methods do benefit the model, the generated words are often meaningless and harms the language modeling performance of the trained character recognition models. Thus, the authors (Park et al., 2023) showed that models trained on the naively generated dataset lacked word-level contextual awareness. The proposed synthetic data generation mixes the data created by up-sampling the tail characters with the original data created from a long-tail character distribution which benefited the model's overall performance. While this exemplifies how using downstream model performance as guidance for data synthesis processes is helpful, it still leaves a lot to be desired as the proposed solution leaves much room for improvement.

Recent works on fine-tuning large language models also show the potential of model guided synthesis in the form of self-refinement (Madaan et al., 2023). In essence, the model that needs to be improved is directly involved in the data generation process, which distinctly sets self-refinement from more academic methods of creating new data. Additionally, as fine-tuning of LLMs expands towards to human preference alignment, various methods have been proposed to create new preference data (Cui et al., 2023; Kim et al., 2023a; Yuan et al., 2024). The generated data all have strong dependencies with the specific model that was used in the synthesis process, yet harbors widespread usage owing to the practical effectiveness of the created data. Thus, model guided data synthesis is becoming the norm in the rapidly evolving world of AI.

3 Disparities in Data Quality Standards: Academic Rigor vs. Industrial Pragmatism

The conceptual divergence between academia and industry regarding "good data" is both nuanced and significant, reflecting deeper methodological and practical differences in how data is curated, maintained, and evolved within these sectors. In academic circles, datasets are often treated as fixed benchmarks. Once a dataset is curated and publicly released, it typically remains unchanged for extended periods. Subsequent versions or enhancements are infrequent, reflecting a culture where the creation of a dataset is seen as a singular scholarly contribution. This approach can lead to notable delays in updates, with some of the most critical datasets in fields like computer vision waiting over a decade for a new iteration, often contingent upon the initiative of a completely different research group (Kim and Park, 2023; Kim et al., 2023b). This scenario underscores a fundamental academic assumption: that the primary value of a dataset lies in its initial creation and its ability to benchmark model performance at a point in time (Recht et al., 2019; Deng et al., 2009).

Contrastingly, the industry adopts a dynamic and iterative perspective on data curation. Datasets are not static but are continuously enhanced through numerous versions, each iteration subjected to rigorous cleansing and modification processes. This relentless pursuit of data quality ensures that datasets remain relevant and reflective of evolving real-world complexities. In this context, good data transcends the traditional academic metrics of task relevance and dataset size. It also encompasses the richness of metadata, which meticulously documents each iteration's changes, providing transparency and traceability essential for ongoing model development and deployment.

Moreover, the industrial emphasis on the iterative cleansing of data introduces additional criteria for evaluating data quality. Factors such as the cost-effectiveness of the cleansing process, the scalability and automation of these procedures, and the generation of comprehensive metadata become pivotal. These criteria highlight the industrial commitment to efficiency, reproducibility, and operational excellence in data management.

The expansion of the notion of good data within the industrial setting to include these operational factors marks a significant departure from academic conventions. While academia primarily focuses on the theoretical and experimental utility of datasets, industry prioritizes practicality, operational efficiency, and adaptability to technological advancements and market demands. This expanded definition encompasses not only the intrinsic qualities of the data but also the processes by which data is curated, maintained, and utilized over time.

This broader, more dynamic conception of good data in industry compared to the relatively static view in academia contributes to the disparity in data quality definitions between these sectors. It highlights a critical need for dialogue and convergence between academic and industrial approaches to data curation, aiming to bridge the gap and foster datasets that are not only scientifically robust but also practically valuable for real-world applications. Such a collaborative approach could enhance the relevance, utility, and impact of data across both research and industry, paving the way for innovations that are deeply informed by both theoretical insight and practical exigencies.

4 Conclusion

The rapid evolution of AI applications within the industrial sector necessitates a reevaluation of our data management strategies. While academic perspectives on data-centric AI offer valuable insights, we propose that the development of industrial AI requires a synthesis of data-centric and model-specific approaches, underpinned by a redefined standard for data quality. This hybrid approach not only leverages the strengths of data-centric methodologies but also integrates model-specific features and approach-specific information, paving the way for innovative methods and enhanced data generation efficacy through model-guided synthesis. Furthermore, an essential component of this industrial paradigm is the development and maintenance of comprehensive metadata that documents each iteration of the data cleansing process. This transparency, coupled with a focus on the cost-effectiveness and automation of data curation, is paramount for setting a new industrial standard for what constitutes good data. In conclusion, the advancement of AI in industrial contexts demands a harmonization of data-centric principles with model-specific insights. By adopting standards that reflect both the importance of data quality and the nuances of model interaction, we can foster the growth of a truly data-centric AI ecosystem, optimized for rapid and efficient industrial development. This shift towards a more nuanced, integrated approach will be critical for realizing the full potential of AI applications in industry.

Broader Impact Statement

This research elucidates the divergent paradigms of "Data-Centric AI" and "Model-Agnostic AI" across academic and industrial landscapes, spotlighting the criticality of data quality standards and their implications for real-world AI applications. By dissecting the nuances between academic and industry data standards, our work sheds light on the existing gaps and underscores the necessity for a harmonized approach towards data quality, paving the way for "Model-Based Data-Centric AI".

References

Inha Cha, Juhyun Oh, Cheul Young Park, Jiyoon Han, and Hwalsuk Lee. Unlocking the tacit knowledge of data work in machine learning. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2023.

Eujeong Choi and Chanjun Park. Dmops: Data management operation and recipes. *arXiv preprint arXiv:2301.01228*, 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, Abhinav Jain, Lokesh Nagalapatti, Sameep Mehta, Sandeep Hans, et al. Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935*, 2021.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling, 2023a.

Damrin Kim, NamHyeok Kim, Chanjun Park, and Harksoo Kim. Transcending traditional boundaries: Leveraging inter-annotator agreement (iaa) for enhancing data management operations (dmops). *arXiv preprint arXiv:2306.14374*, 2023b.

NamHyeok Kim and Chanjun Park. Inter-annotator agreement in the wild: Uncovering its emerging roles and considerations in real-world scenarios. *arXiv preprint arXiv:2306.14373*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

Tarun Narayanan, Ajay Krishnan, Anirudh Koul, and Siddha Ganju. Curator: Creating large-scale curated labelled datasets using self-supervised learning. *arXiv preprint arXiv:2212.14099*, 2022.

Sunghyun Park, Sunghyo Chung, Jungsoo Lee, and Jaegul Choo. Improving scene text recognition for character-level long-tailed distribution. *arXiv preprint arXiv:2304.08592*, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.