ON GENERATED VS COLLECTED DATA

Levent Sagun, Elvis Dohmatob, Kartik Ahuja, Julia Kempe
FAIR Paris
{leventsagun}@meta.com

Recent developments in generative AI models heavily rely on the abundance of high-quality data which may not be readily available. In particular, in applications when data is scarce or costly, practitioners started using these very same models for data generation and evaluation. However, such self-generated data has been shown to be of unreliable quality and may be prone to hallucinations [9]. And these effects are amplified with repeat applications [1]. Furthermore, evaluations relying on trained models may suffer from inaccuracies especially on topics that are at the margins of the datasets that the models are originally trained with. Global inclusion necessitates intentional inclusion of those in the margins, of those that are not represented by dominant datasets.

In this short note, we outline two theoretical arguments as to why self-reliance may only work for typical or simple samples of the training data. While self-reliance might be useful for some part of the data, this might come at the expense of poorer performance for the part of data that are more complex or rare. As inclusive models are those that can't ignore the parts of data in the margins, we argue for explicit human intervention via targeted data collection. To go beyond the statistical limitations of existing models, human evaluation and participatory methods shall be used in tandem.

Our first argument considers the case of out-of-distribution generalization. The goal here is to train a classifier on data from a mixture of environments that can perform well on new environments. The new environments and the seen environments share a fundamental invariance – the label generation function stays the same. If the training environments are not sufficiently diverse, then it is typically impossible to learn this invariance [6]. Consider a model trained on such training environments with the aim of generating more data for augmentation, hoping to improve out-of-distribution generalization. This model can be interpreted as creating new environments that are equivalent to a mixture of the seen training environments. These new artificial environments do not add any real diversity in ways that would permit the learning of the invariance.

Our second argument is through the lens of bulk-tail separation of the data distribution. In cases when data is heavy-tailed, as it is in many real-world instances, self-reliant data generation may omit the tail during the process. Thus iterative re-generation of data by AI leads to poorer and poorer performance on rare parts of data, leading to model collapse [7, 3, 4]. In particular, [4] shows that increasing dataset size leads to power scaling laws. And when data is generated by AI models, the undersampling of tails leads to stalled, tapered-off error scaling curve because AI-generated data only captures the bulk of the statistically most important aspects of real-data up to a certain point k, but omits the tail. This leads to a modification of the classical neural scaling laws [5], where k is now an additional scalable parameter, alongside the usual sample size T and model size N.

Both of the above arguments hint in the same direction: In order to preserve performance and prevent successive worsening on marginalized or non-typical data, diverse tail data needs to be acquired. This doesn't mean that generated data has no use. Indeed, in some special cases, when the training environments are sufficiently diverse but some are far more represented in the data, the generation of data to augment and balance these environments can help towards learning invariances [2]. Another example comes from finite environments where the reward function is explicitly accessible to the model and therefore there is no need for costly human annotation or unreliable AI based evaluation. In this case, models can self-explore the environment and be trained on self-generated data, a strategy proven to be very useful on out-of-distribution generalization, for example, in games [8].

However useful self-generated data may be in certain contexts, we should also consider that many recent applications of generative AI produce text, image, speech, and video. And such data are vastly more complex and heavy tailed that the search space is intractable to ensure complete coverage by AI generated data. Furthermore appropriate evaluation of the generated outputs are very costly and hard to automatize especially for the tail and environments that are less observed during training.

Based on this discussion of potential pitfalls and benefits of generated data, we argue in support of context aware data collection efforts towards tails of data distribution to complement automated data generation in order to be truly globally inclusive.

REFERENCES

[1] Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023.

[2] Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? 2023.

[3] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. *arXiv preprint arXiv:2402.07712*, 2024.

[4] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

[5] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[6] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[7] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. 2023.

[8] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529 (7587):484–489, 2016.

[9] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.