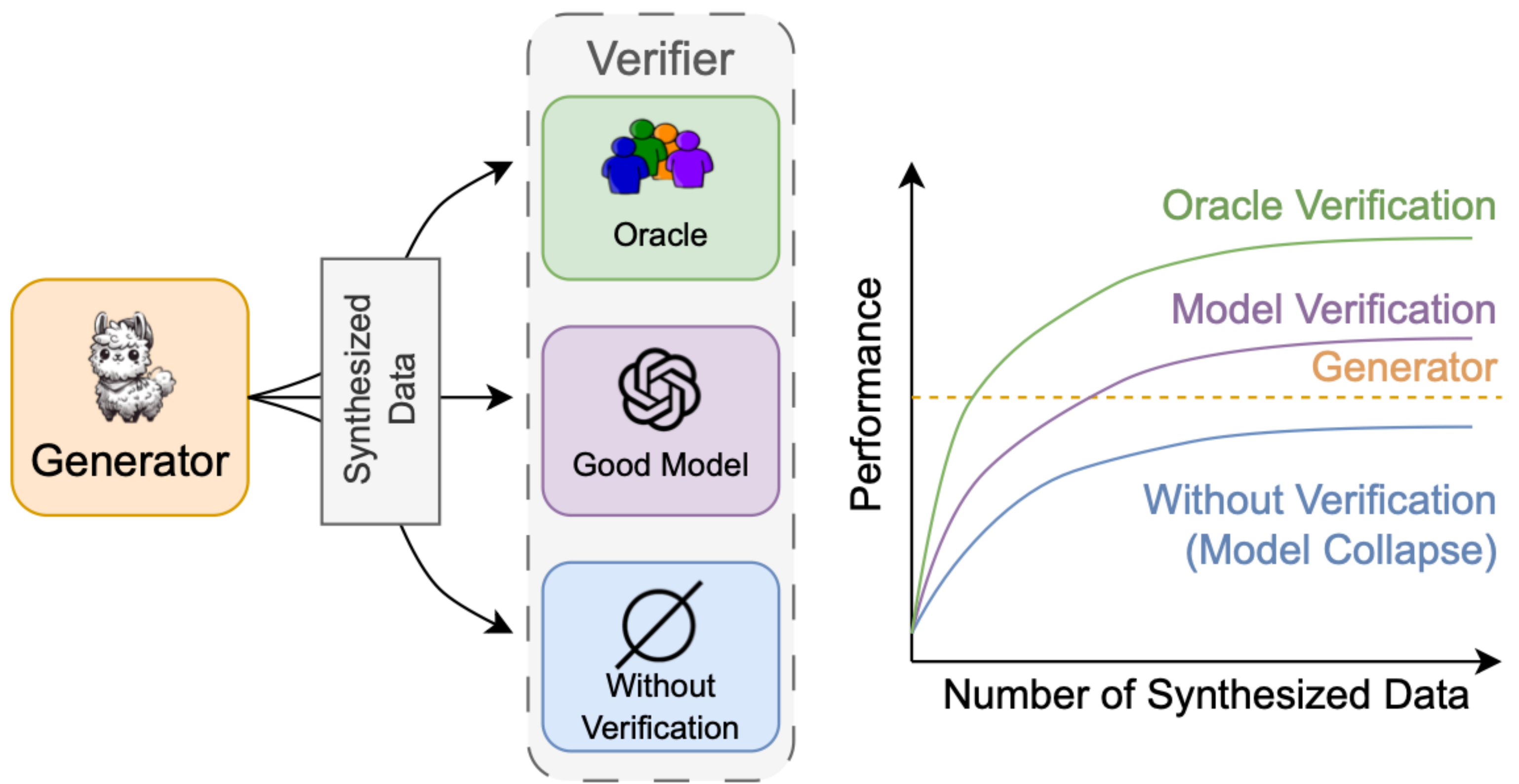
Meta
NYU
PEKING UNIVERSITY

Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification

Yunzhen Feng* Elvis Dohmatob*
Pu Yang* Francois Charton
Julia Kempe

TL;DR. AI-generated data could pose risks like model collapse and changes of scaling laws. We find that verifying synthetic data is key to its utility. Our research, through theory and experiments, characterize conditions for (weak) verifiers to sufficiently improve synthetic data to train models that exceed the performance of the original generator. We provide a metric for verifier quality.

Synthetic data generated by contemporary large-scale models present significant opportunities. However, concerns arise regarding potential risks associated with model collapse. When the training set incorporates lots of synthetic data, issues such as performance degradation, training instability, and loss of scaling may occur. We show that scaling up with synthetic data requires verification in data curation.



Warmup: what is in the generated data? Case study of eigenvalue prediction

Transformer trained to predict eigenvalues of 5 x 5 matrices given all the entries. y' as the prediction.

Evaluate accuracy with $1_{\{\frac{|y-y'|}{|y|}<\tau\}}$, where s is a threshold scaler.

200,000 Data points (x, y) → Trans-former (Accuracy 67% (τ= 2%)) → generate12M Data points (x, y') → Verifier → pass 1% threshold (about 20%) Keep! → Trans-former: Accuracy 95%! (τ = 2%)

Verifier → fail 1% threshold (about 80%) Toss!

Verification (reinforcement) allows not only to avoid model collapse, but to bootstrap data!

In hindsight, this aligns with R1-zero: it trades computation for high-quality, verified synthetic data with rule-based rewards.

	Tolerance τ 2%	1%	0.5%
Data Selection 2%	72.1	20.2	2.3
Label Selection Beam 50	**84.0**	**33.4**	**4.9**
Label Selection Beam 25	79.9	28.7	4.1
Label Selection Beam 10	73.9	22.7	2.9
Label Selection Beam 5	69.1	19.0	2.3
Greedy w/o selection	60.5	14.5	1.7
Synthesized Generator	66.9	20.2	2.4

	Verify all beams			Verify the best beam		
Tolerance τ	2%	1%	0.5%	2%	1%	0.5%
Beam 50	90.4	60.4	22.9	65.9	19.2	2.4
Beam 35	89.2	56.9	19.8	66.0	19.2	2.4
Beam 25	88.0	53.2	16.8	66.1	19.3	2.4
Beam 10	83.7	43.1	10.5	66.2	19.5	2.5
Beam 5	79.3	34.9	7.1	66.5	19.7	2.4
Greedy	66.9	20.2	2.4	66.9	20.2	2.4

To select: use verifier vs use model itself

Left:

Model Collapse: Greedy w/o Selection vs Synthesized Generator.

Reinforcement: Both label and data selection significantly enhance outcomes.

Right (*fun fact*):

The model itself is unable to distinguish high-quality outputs -> we need an external verifier!

Theoretical Metric

Suppose y' is the generated part. $p = P(y' \neq y)$ is the accuracy.

q denotes if the sample is kept after the verification.

Define $\phi_k = P(q = 1|y' = k, y = k), \psi_{kl} = P(q = 1|y' = l, y = k)$, as verification-related constants.

Suppose symmetric pruning with two classes (correct/false), $\phi_0 = \phi_1 = \phi$ (True Positive), $\psi_{01} = \psi_{10} = \psi$ (True Negative).

$$\text{Define } p_* = 1/(1+\frac{\psi}{\phi}).$$

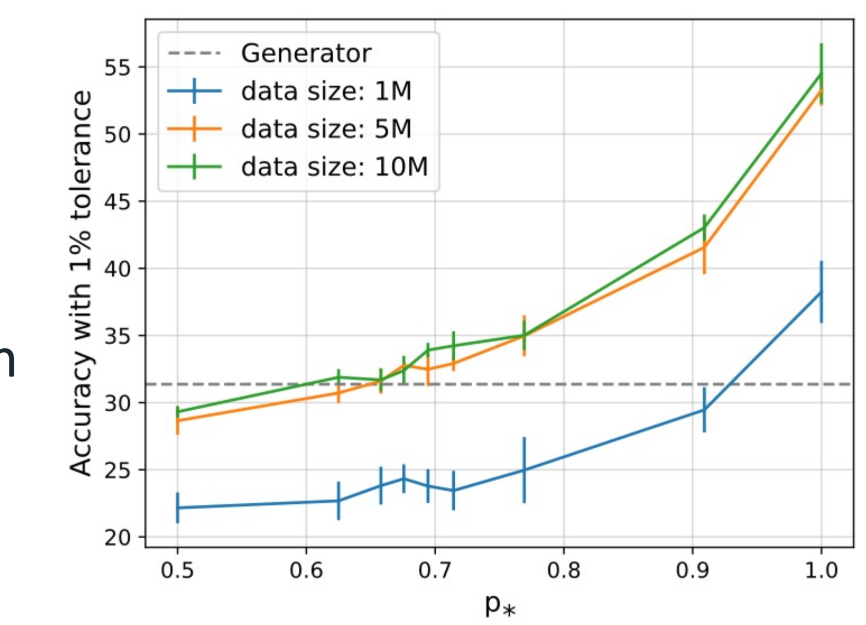
Theorem: Assume data from a Gaussian mixture, **linear** models, **linear** verifier: in the high-dimensional limit, the final model trained with curated data achieve Bayes-optimal performance when the accuracy of generator (1- p) is larger than p^*.

Remarks:

- p_* depends on the generator, the verifier, and the ground truth.
- We could use p^* as a metric for the current verifier, if it is strong enough to curate synthetic data.
- $\boldsymbol{p_*}$ **does not simply follow verifier accuracy**. A model that performs better in terms of classification accuracy can, counterintuitively, be a weaker verifier.

Experiments

p_* successfully predicts when the trained model outperforms the generator. In the eigenvalue prediction setup.



Llama 2 on News Summarization

- Finetune Llama 2 on English News Summarization with XLSUM.
- The generator is trained with 25% data.
- Generate and verifier-filter synthetic data on the rest.
- Use 1) the model itself (ppl), 2) finetuned Llama 3 (ppl), 3) ground truth (string match, ROUGE) as verifier.

1) Model collapse: Random Selection vs Generator.
2) Oracle selection: surpasses model collapse and outperforms training with real labels (full origin).
3) Self selection: have improvement. Similar to LLM as a judge using the generator.
4) Llama 3 selection: Ineffective; though the Llama 3 model has higher performance. A better model is not necessarily a better verifier (echo the remark).