

Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification

Y. Feng, E. Dohmatob, P. Yang, F. Charton, and J. Kempe
ICLR 2025

背景：合成データの混入と利用

インターネット上の合成データの蓄積

- 生成された文章や画像が事前学習データへ混入
 - 混入後に元データと合成データを分離することは困難

合成データの学習利用

- 質問応答・コード生成・数学などで正解データを合成
 - 人手による作成より低コストで量の拡大が容易

背景：合成データでの学習

前世代の出力による反復学習

- 数世代でモデルの分布が元の分布から乖離
 - 出力は元データと無関係となり,文章や画像で同様の退行

スケーリング則の変化

- 合成データではデータ量を増やしても性能向上に限界がある

既存の合成データへの対処

- 実データに近い合成データを優先して利用
- 合成データを補正関数で書き換えて利用

目的・アプローチ

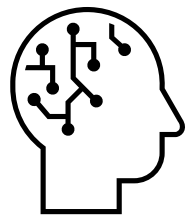
本論文が立てる問い

- 破棄や修正以外で合成データを利用できないか

アプローチ

- 合成データを検証器に通し残ったもので学習

→ 設計の対象は生成器の強化ではなく検証器の質



生成器

元データで学習した
1つ前のモデル



合成データ

入力に対する答え

正誤の混在



検証器

答えの良否のみを判定し、
通すか捨てるかを決定

予備実験：生成器の出力の質

タスク：5×5 実対称行列の固有値の予測

Transformerモデルによるbeam searchで生成

真値との相対誤差1%以内の出力を正解とみなす

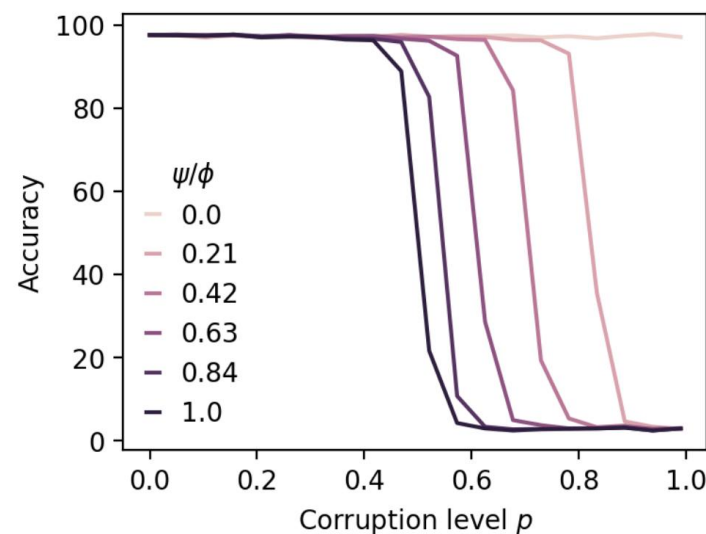
同時に保持する候補の数	候補の中の最良解(正解率)	モデル自身が選んだ解(正解率)
1（候補を広げない）	20.2	20.2
5	34.9	19.7
10	43.1	19.5
25	53.2	19.3
50	60.4	19.2

→ 候補を拡大してもモデル自身の選択で性能に限界がある

理論：崩壊点

崩壊点 p^*

- 誤り率 p がこの値を超えると学習後の性能が大きく低下
- ϕ は正しいデータ, ψ は誤ったデータを通す割合
- $p^* = 1 / (1 + \psi/\phi)$
- 選別なしで $p^* = 0.5$, 完全な検証器で 1



横軸は合成ラベルの誤り率 p , 縦軸は学習後の精度(%)

ψ/ϕ が小さい検証器ほど p^* は右へ動く

p^* を超えた時点での精度の**反転**

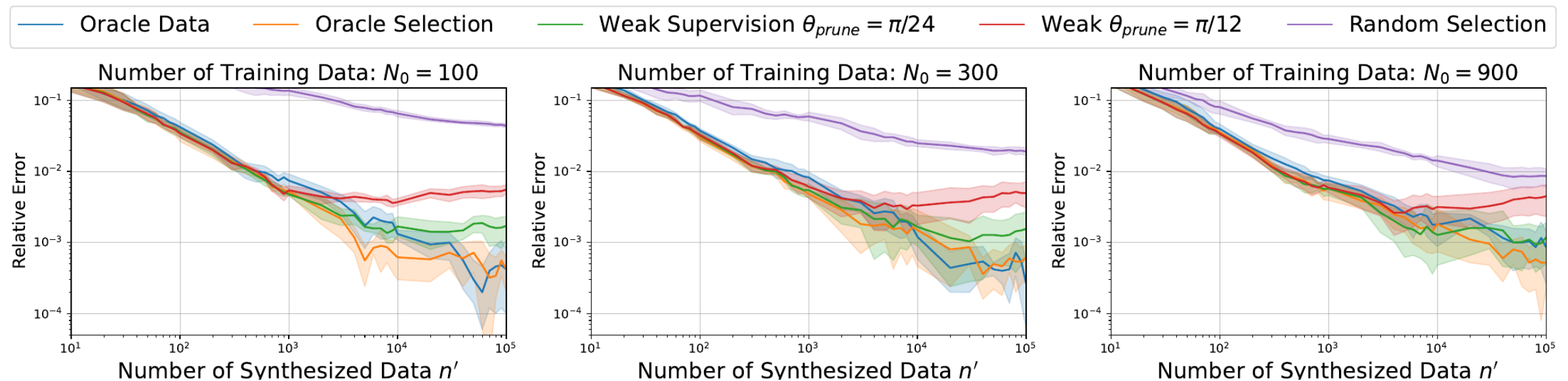


ある誤り率を境に一挙に生じる性能低下

シミュレーション：有限のデータ量

データ量が有限の場合の挙動

- 理論と同じガウス混合分布での,データ量を変えた学習



横軸は選別後に学習へ用いた合成データ数,縦軸は最適な精度からの誤差

完全な検証器による選別は,正しいラベルのみでの学習とほぼ一致

弱い検証器では,ある量を超えると誤差が下げ止まり

→ 適切な検証器であれば合成データでも有効な学習が可能

実験：3つの設定

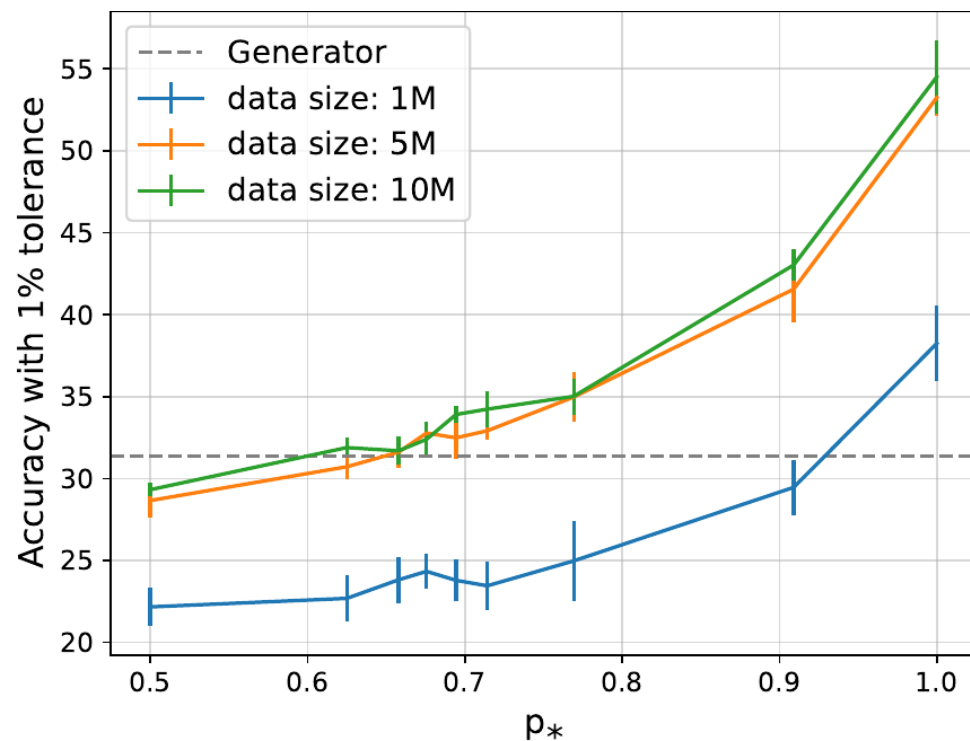
設定	扱うデータ	モデル	検証器	評価指標
シミュレーション	ガウス混合	線形モデル	線形モデル	分類の正解率
固有値の予測	合成した行列	Transformer	正解と照合	許容誤差つき 正解率
ニュース要約	ニュース記事	LLM	LLM	単語の重なり

→理論の仮定から実タスクへ,段階的に条件を緩めて検証

実験1：固有値の予測

検証器の質と学習データ量の操作

- 誤った答えを一定の確率で通す検証器を用意
- p^* を 0.5 から 1 まで変化させる



選別なし ($p^*=0.5$) は生成器よりも**低精度**

完全な検証器 ($p^*=1$) では正解率がほぼ**倍増**

同一のデータ量では p^* が高いほど高精度

相対誤差 1% 以内の正解率(%). 破線は生成器

➡ 性能を決めるのはデータ量ではなく検証器の質

実験2：実験設定

XLSUM の英語ニュース要約

- 学習データの 12.5% で微調整した Llama-2-7B を生成器に設定
- 生成器による全記事の要約生成と、人手の要約との重なりでの評価

4つの選び方の比較

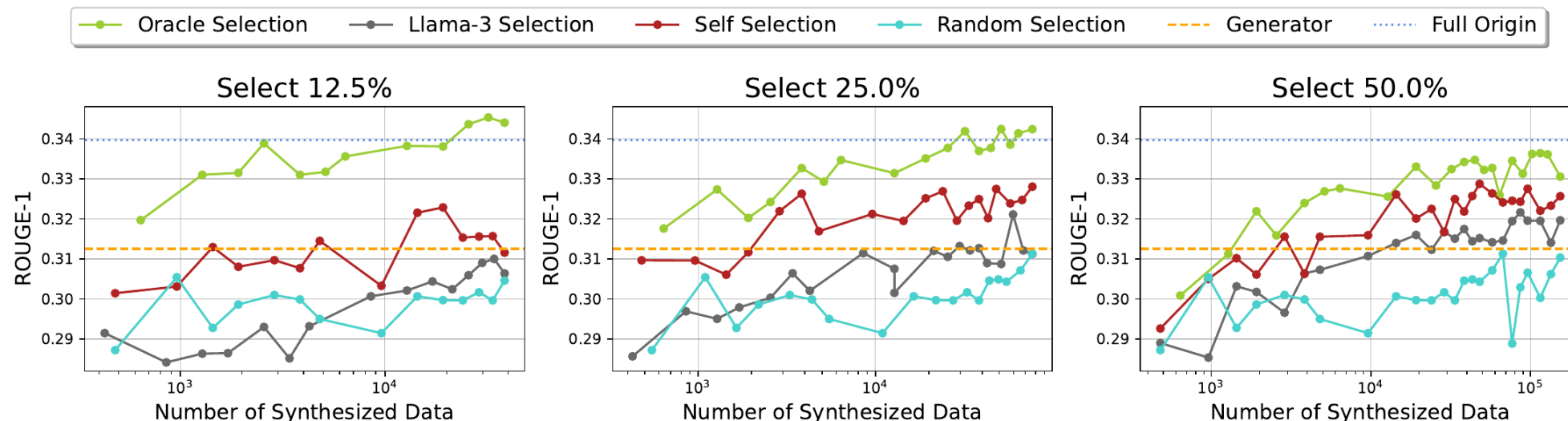
- 正解との一致・別のモデル・生成器自身・選別なし

→ 生成器が学習に用いない記事からの選別

実験2：選別の効果

選別する割合の比較：12.5%・25%・50%

- 縦軸は人手の要約との単語の重なり,横軸は合成データ数



選別なしは,生成器より**低精度**

正解での選別は,全データを人手ラベルで学習したものより**高精度**

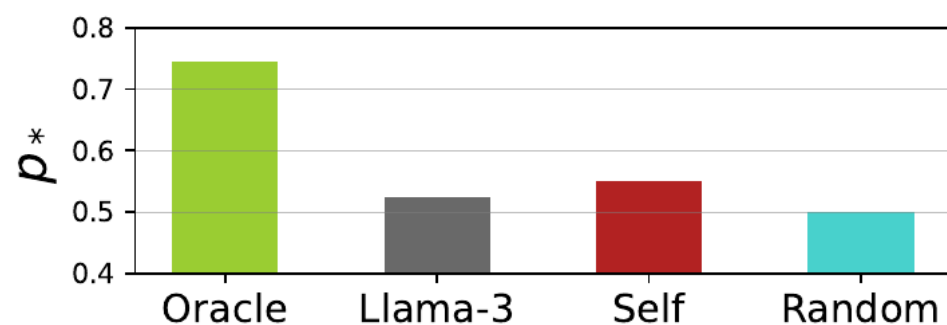


合成データの増量より選別の方が効果的

実験2：検証器の選び方

要約で優れるのは生成器自身による選別

- より良い要約を書ける Llama-3 による選別でも,選別なしと同水準



選別方法ごとの p* (選別率 12.5%)

合成データの生成元 Llama-2 と, Llama-3 の尺度との低い相関

測定した p* の順序も自己選別が優位

→ 検証器の選択基準は強さではなく生成器との相性

考察：自己改善は可能か

固有値の予測では自己選別は機能しない

- 入力ごとに良い答えを選んでも,選ばれた答えの分布は不変

要約では生成器を上回る自己選別

- 生成器が学習に用いた記事より広い集合からの,良い組の選択
 - 入力側の分布の変化による新しい情報の追加

→ 自己改善の成否を決めるのは,入力の分布の広がり

結論

合成データには学習に十分な情報が残っている

- 適切な選別により正しいラベルでの学習と同水準

ある誤り率を境に一挙に崩壊が生じる

- 崩壊点 p^* を決めるのは検証器が正誤を選り分ける精度

学習の実施前に p^* が測定可能

- 学習前に検証器を比較することが可能

利用上の示唆

- 生成器の強化に先立つ, 利用可能な検証器の有無の確認