

How to Synthesize Text Data without Model Collapse?

Xuekai Zhu^{1,2} Daixuan Cheng² Hengli Li^{2,3} Kaiyan Zhang⁴ Ermo Hua⁴ Xingtai Lv⁴ Ning Ding⁴
Zhouhan Lin^{†1,5} Zilong Zheng^{†2} Bowen Zhou^{†4,5}

ICML 2025(International Conference on Machine Learning)

発表：藤原 颯太

背景：合成データによるモデル崩壊

LLMの普及による合成データの広がり

- 合成データは多くのインターネットコンテンツに含まれる
 - ソーシャルメディア
 - ブログ
 - オープンソースリポジトリ
- 合成データの学習は不可避になる

モデル崩壊の危険性

- 再帰的学習による性能低下
- スケーリング則の打切り
 - 合成データを増やしてもモデルの性能は向上しない

課題

非反復的モデル崩壊

- 異なるモデル由来の合成データでの実験
- 先行研究では明らかになっていない

モデル崩壊を防ぐデータ合成

- 人間データは有限で高コスト
- 合成データには人間データを強化できる可能性がある

問題の実証：実験設定

使用モデル

- GPT-2 Small
- OLMo-237M

使用データ

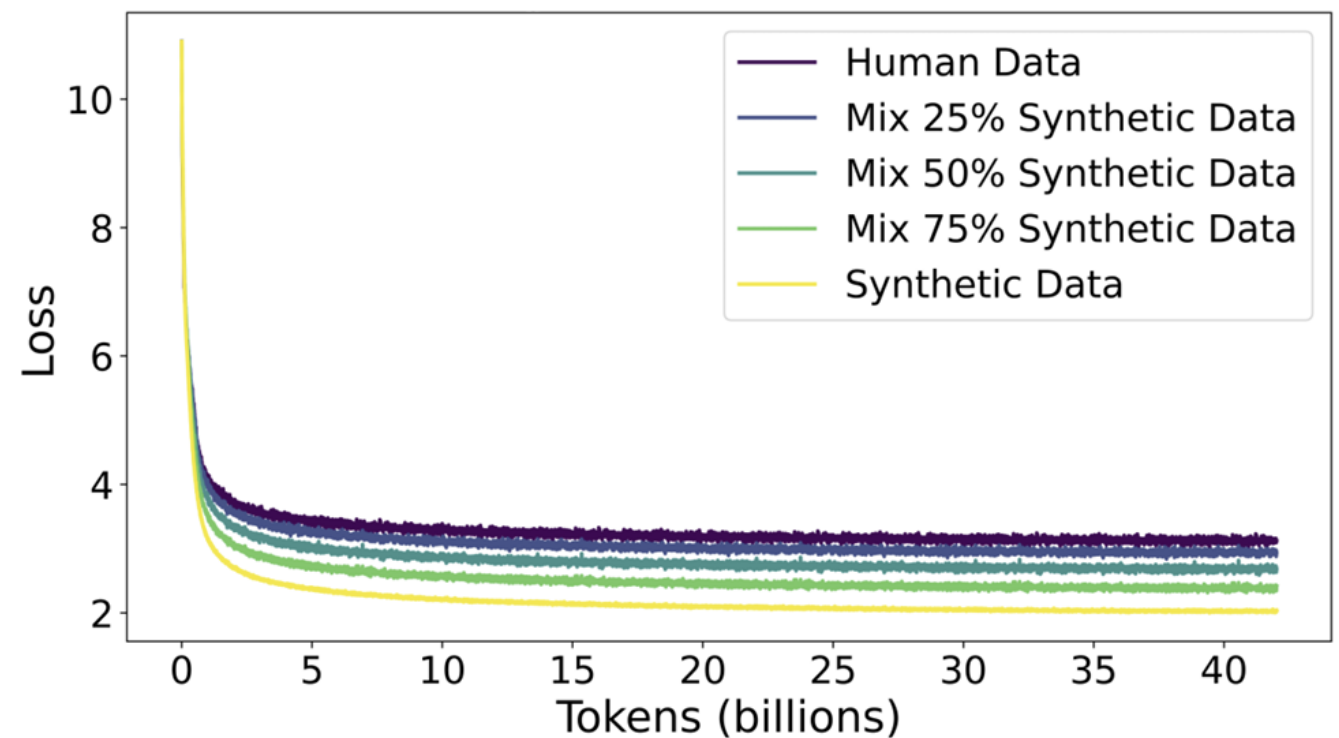
- 人間データ
 - Dolma(Wikipediaなど人間が書いたデータを集めたデータセット)
- 合成データ
 - Cosmopedia(Mixtralを用いて作成されたデータセット)

問題の実証：結果

合成データ量と予測精度

- 合成データの割合を増加させるほどモデルの予測精度は低下
- 訓練ロスは合成データのみの場合が最も低い
 - 訓練データへの過適合

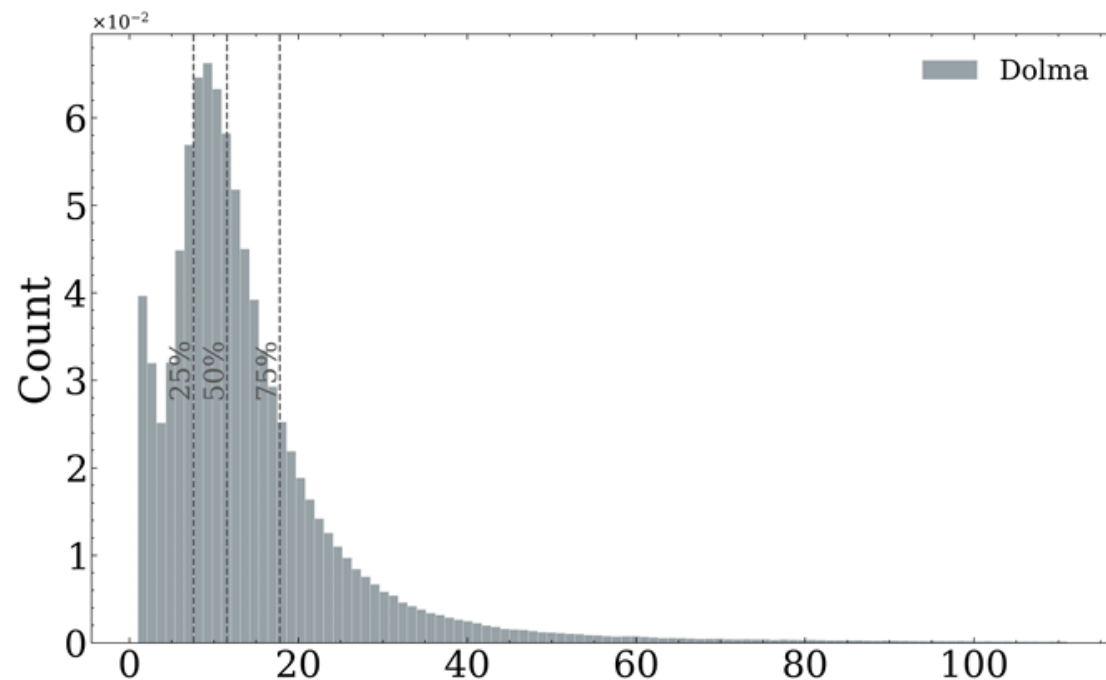
合成データの割合	平均PPL
0%	約22
25%	約24
50%	約25
75%	約28
100%	約52



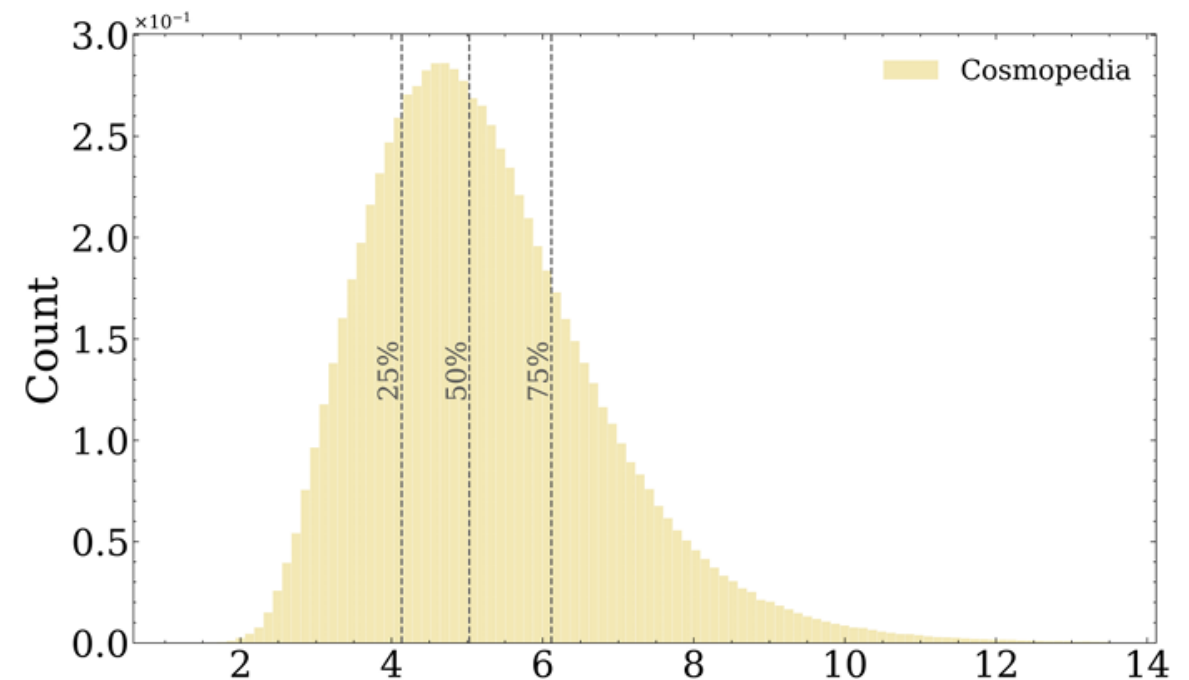
問題の原因説明①

カバレッジの狭窄

- 合成データと人間データは分布が異なる
- 合成データは珍しい表現をカバーできない
- 人間データのようなロングテールを生み出せない



A. Human Data PPL $\in [0, 100+]$

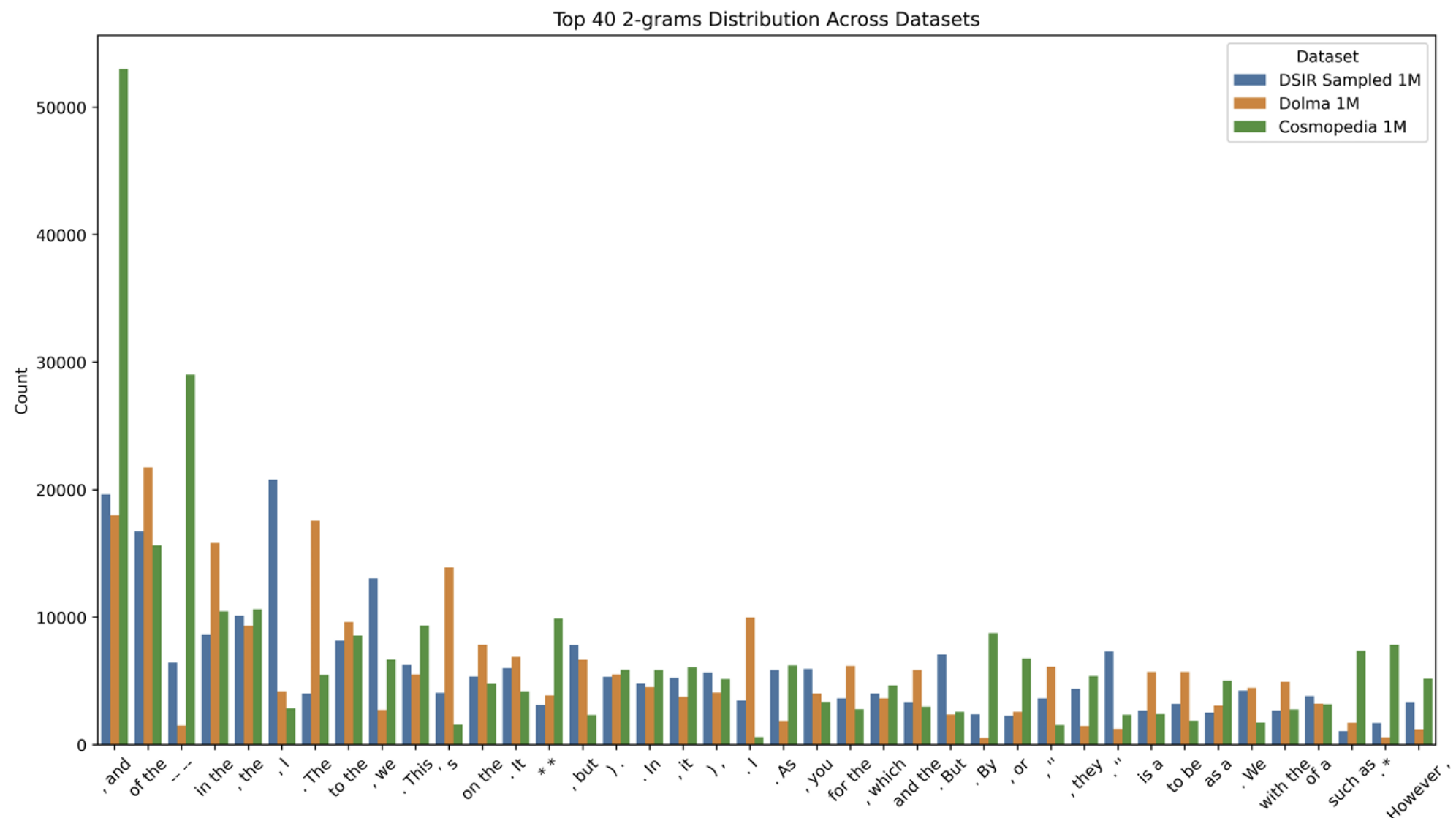


B. Synthetic Data PPL $\in [0, 14]$

問題の原因説明②

N-gram特徴の過集中

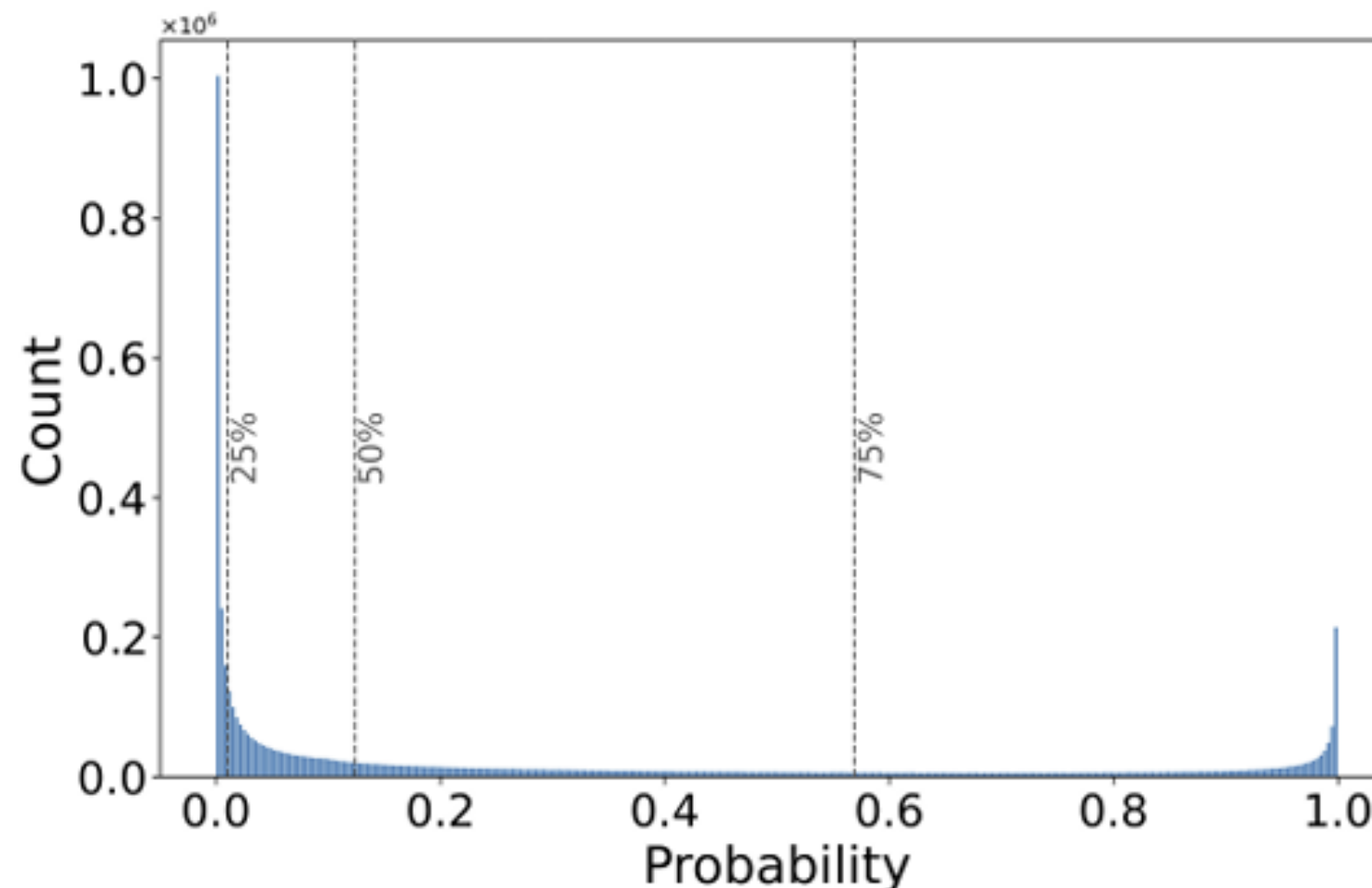
- 特定のバイグラムが高頻度で出現する
- AIは同じような言い回しを使いがち



提案手法：ToEdit①

起点：人間データのU字型分布

- 非常に簡単なトークンと難しいトークンが多い
- 中間の難度のトークンが少ない



提案手法：ToEdit②

手法：トークンレベル編集

- 人間データの予測が容易なトークンをLLMで置き換える
- 頻出の表現を置換することで表現を多様化する
- 人間データのロングテールを保持
- N-gramの過集中を防止

提案手法：ToEdit③

理論的証明

- テスト誤差の上界は有限で反復回数に依存しない
- 反復学習を繰り返しても一定レベル以上低下しない

- 従来手法のテスト誤差

$$E_{test}(\hat{w}_{n+1}) = \frac{\sigma^2 d}{T - d - 1} \times n$$

- ToEditでのテスト誤差

$$E_{test}(\hat{w}_{n+1}) \leq \frac{2\sigma^2 d}{T - d - 1}$$

結論

- | 反復的でない学習でも合成データを用いるとモデル崩壊は発生する
- | モデル崩壊の原因は「カバレッジの狭窄」と「N-gramの過集中」であり,データ選択でも修正できない
- | モデル崩壊を発生させないトークンレベル編集ToEditは理論的にも実験的にも有効である

所感

半合成データのアプローチについて

- 既知の学習パターンを置換して人間データの質を高めるというのが面白い
- 合成データを学習に使えるようになるというわけではないのでデータ量の問題や合成データによる汚染はやはり防げない
- 汚染済みのモデルの再生や合成データで学習を進める手法についてはまだ課題が残っていると感じた