How to Synthesize Text Data without Model Collapse?

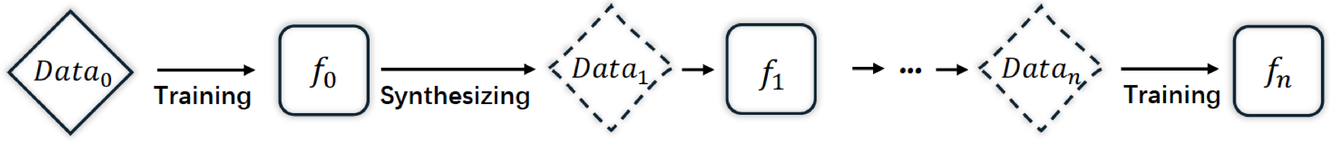
Xuekai Zhu · Daixuan Cheng · Hengli Li · Kaiyan Zhang · Ermo Hua · Xingtai Lv · Ning Ding · Zhouhan Lin · Zilong Zheng · Bowen Zhou

Code

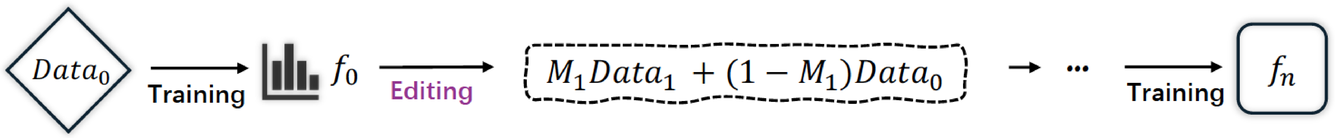
Background: Future GPT-{n} models will inevitably be trained on a blend of synthetic and human-produced data.

Model collapse in synthetic data indicates that iterative training on self-generated data leads to a gradual decline in performance.

① Model Collapse Setting ⇢ $\boldsymbol{E_{test} = \frac{\sigma^2 d}{T-d-1} \times n}$



② Token-Level Editing ⇢ $\boldsymbol{E_{test} \leq \frac{\sigma^2 d}{T-d-1} \times 2}$ ⇢ Avoiding Model Collapse



Trained Model f_i

Source Real Data: $Data_0$

Synthetic Data: $Data_{>0}$

Iterations $i \in \{1, \ldots, n\}$

Test Error E_{test}

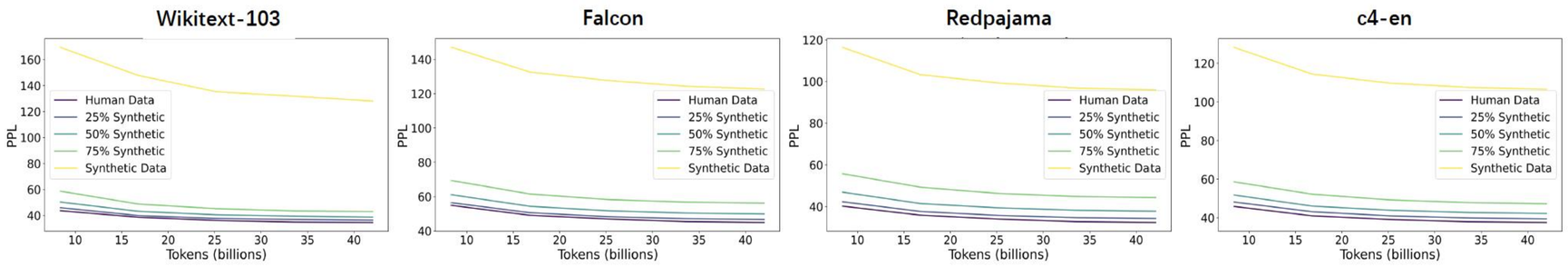
Data Size T

Input Dimensions d

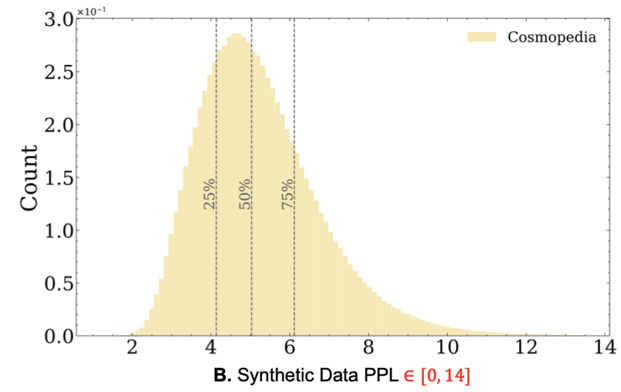
Label Noise Scalar σ

Editing Operation Matrix M_i

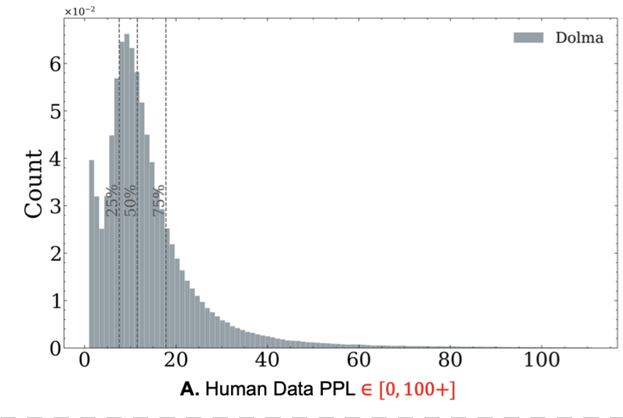
Non-iterative model collapse



Vanishing of the Long Tail



B. Synthetic Data PPL ∈ [0, 14]



A. Human Data PPL ∈ [0, 100+]