

A Tale of Tails: Model Collapse as a Change of Scaling Laws

Elvis Dohmatob^{*1} Yunzhen Feng^{*2} Pu Yang³ Francois Charton¹ Julia Kempe^{2,4,1}

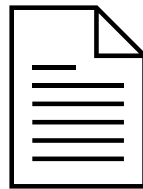
ICML 2024

発表：藤原 颯太

背景：スケーリング則

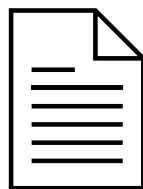
LLMのテスト誤差はデータ量の増加とともに減少

少量のデータで学習



テスト誤差：大

大量のデータで学習



テスト誤差：小

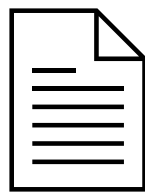


学習に用いるデータが増加するほど
LLMは賢くなる

背景：スケーリング則

LLMのテスト誤差はデータ量の増加とともに減少

少量のデータで学習



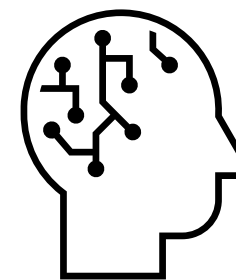
テスト誤差：大

大量のデータで学習



テスト誤差：小

合成データで学習

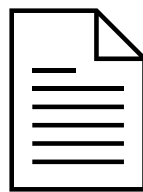


テスト誤差：？

背景：スケーリング則

LLMのテスト誤差はデータ量の増加とともに減少

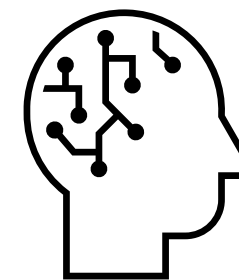
少量のデータで学習



合成データで学習すると
テスト誤差はどうなる？

テスト誤差：小

合成データで学習



テスト誤差：？

背景：合成データとモデル崩壊

合成データの問題点

- 合成データは学習データの分布を完全には再現できない
- 希少な事例が失われ多様性が低下する

モデル崩壊

- 合成データを繰り返し学習に用いることで発生
- モデルの分布が実データから乖離

研究目的・アプローチ

研究目的

- 合成データを用いた際のスケーリング則の変化を調査
- モデル崩壊をスケーリング則の変化として定式化

アプローチ

- 簡易的なモデルを用いて理論的に証明する
- 実験により理論の妥当性を検証

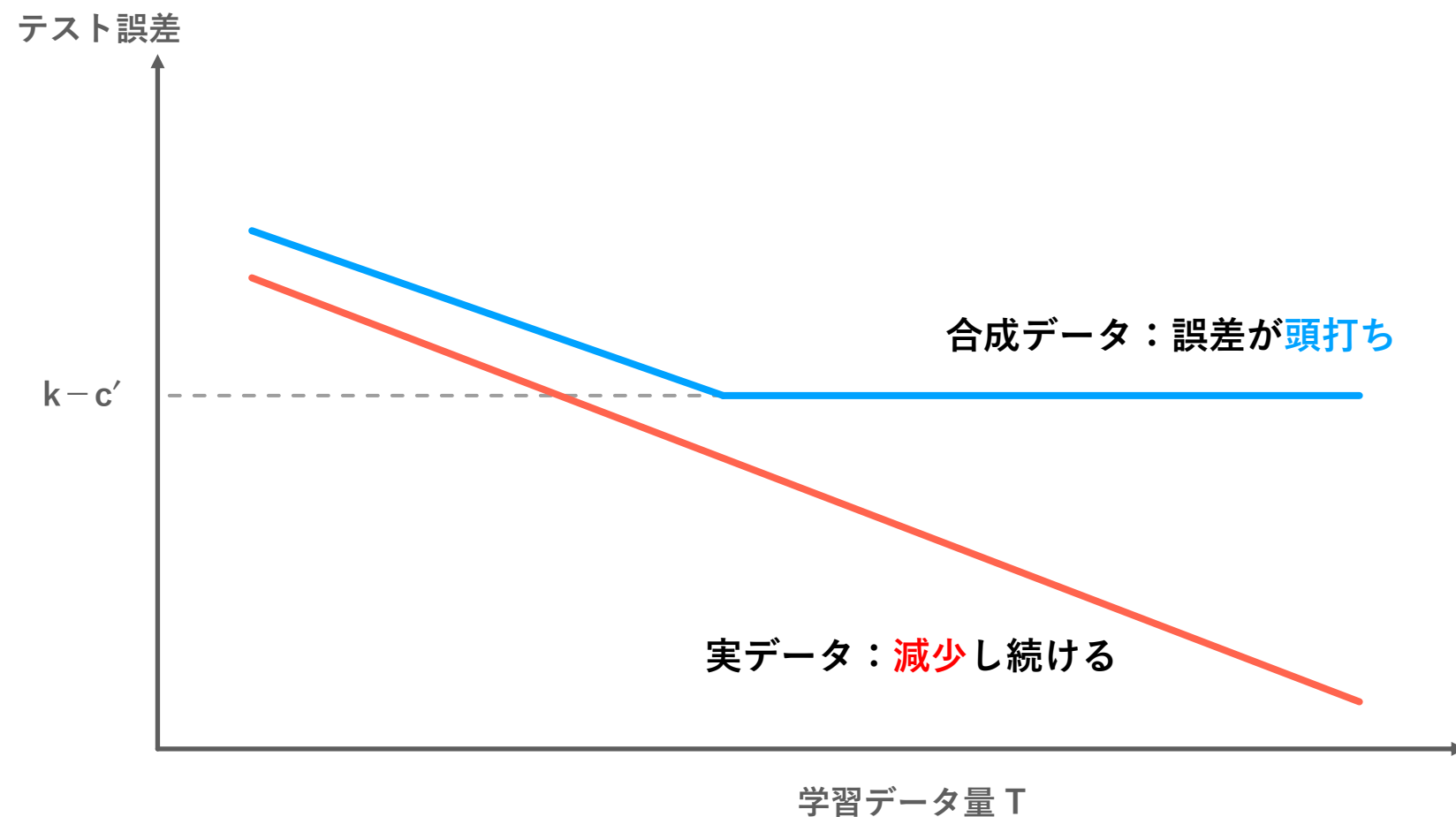
理論①：二重スケージング則

合成データでの学習は誤差が頭打ちに

$$E_{test} \asymp T^{-c} + k^{-c'}$$

k : テールの切断ランク

- T をいくら増やしても $k^{-c'}$ 以下には下がらない



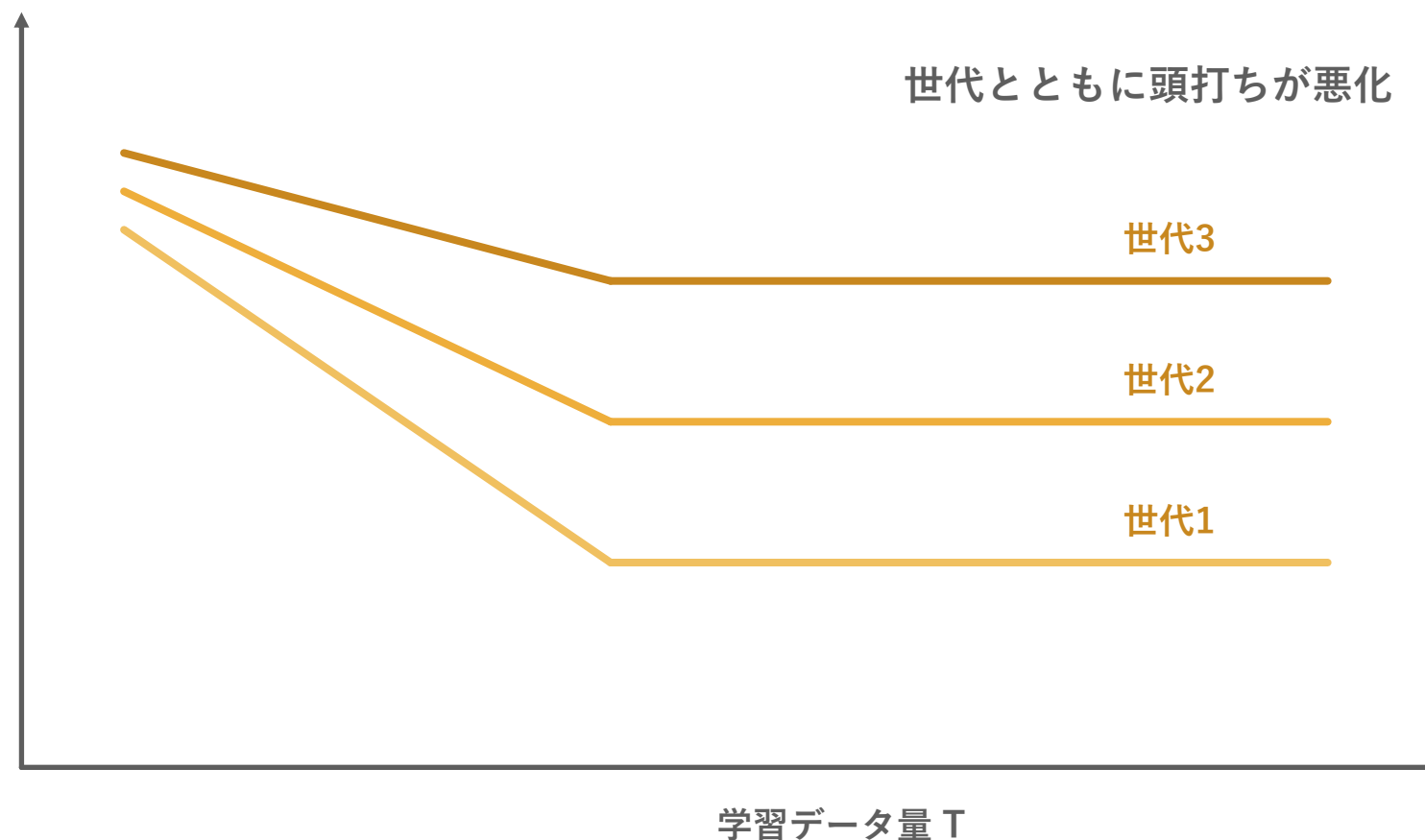
理論②：世代を重ねたモデル崩壊

世代数 n に比例して劣化が蓄積

$$E_{test} = E_{clean} + n \times \text{新しい誤差項}$$

- 頭打ちの項が $k^{-c'}$ から $nk^{-c'}$ へ増大
- 世代が進むほどWebが汚染され，学習が困難になる

テスト誤差

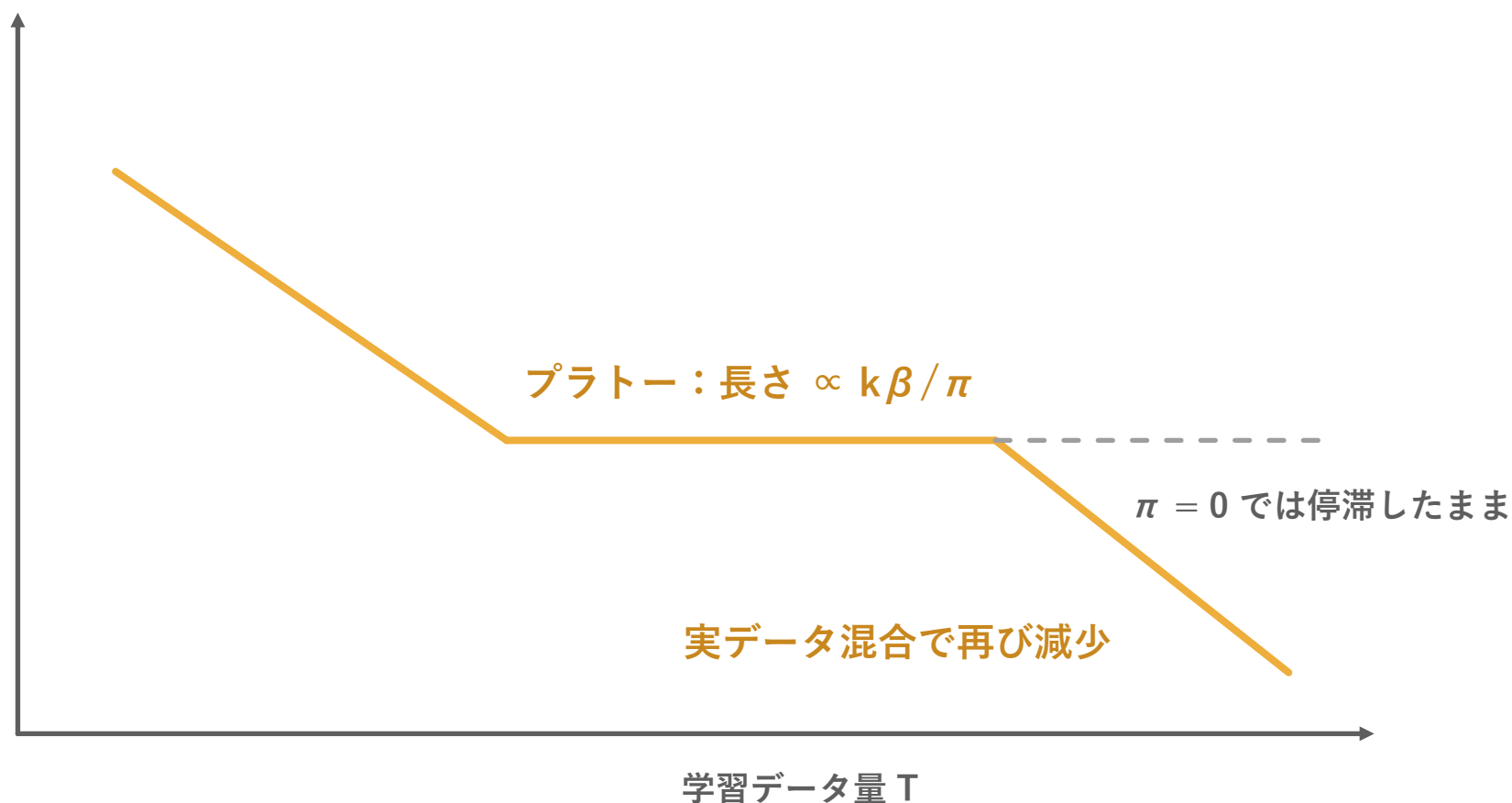


理論③：実データ混合による緩和

実データを割合 π だけ混ぜると崩壊を緩和できる

- Grokking：プラトーと呼ばれる停滞の後に誤差が再び減少
プラトー長 $\propto k^6/\pi$
- $\pi = 0$ では永遠に停滞

テスト誤差



実験設定

GCD Transformer : 算術タスク

- 2整数の最大公約数を予測
- 正解は一意に定まる
- 300Mサンプル学習後のモデルを生成器に利用

Llama2-7B : 大規模LLM

- Wikitext-103 で LoRA 微調整を反復
- プロンプトの続きを $\text{top-p} = 0.9$ で生成し次世代の学習データに

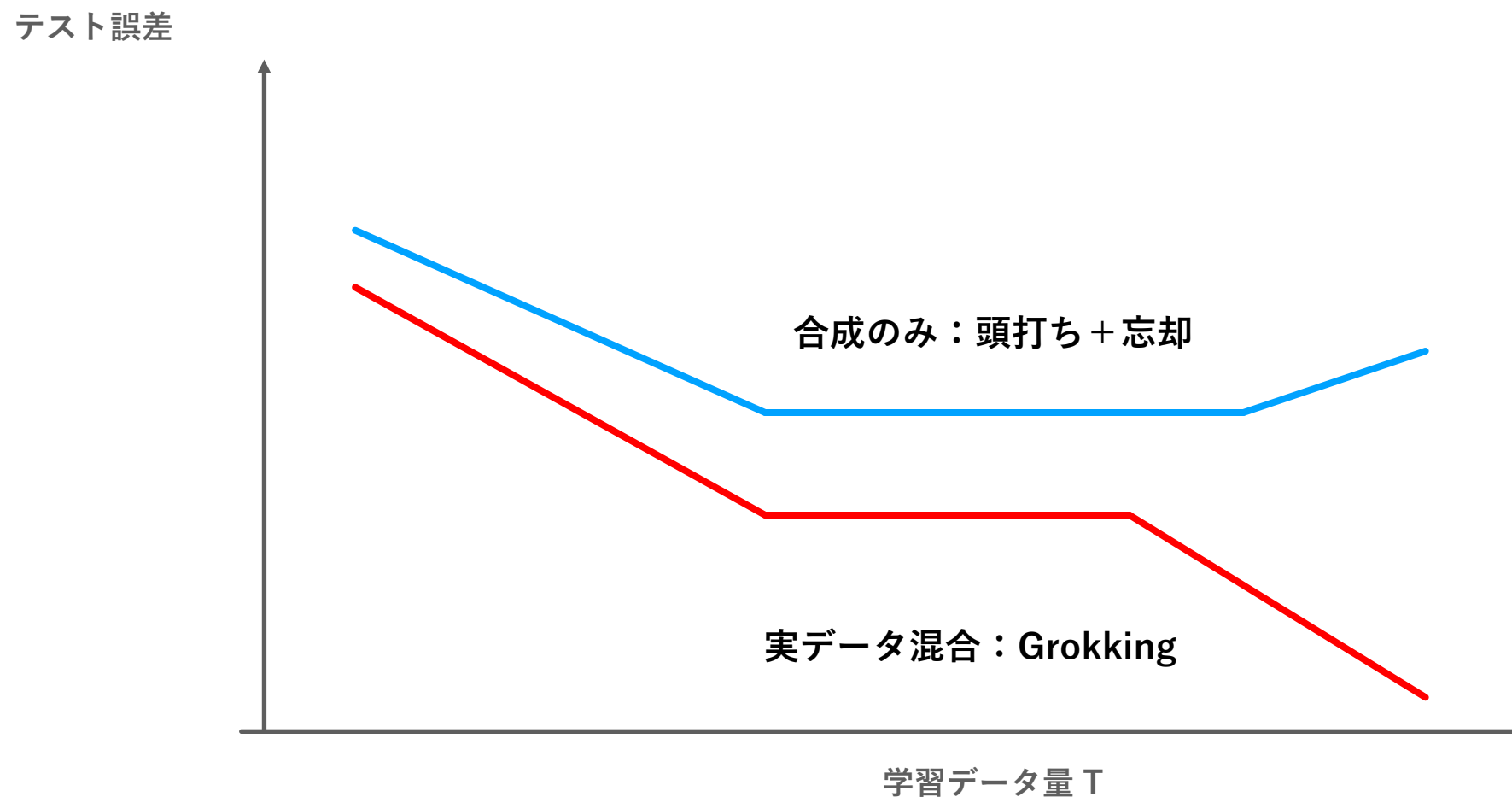


この手順を世代ごとに反復

実験結果①：GCD Transformer

理論の予測とおりの挙動を確認

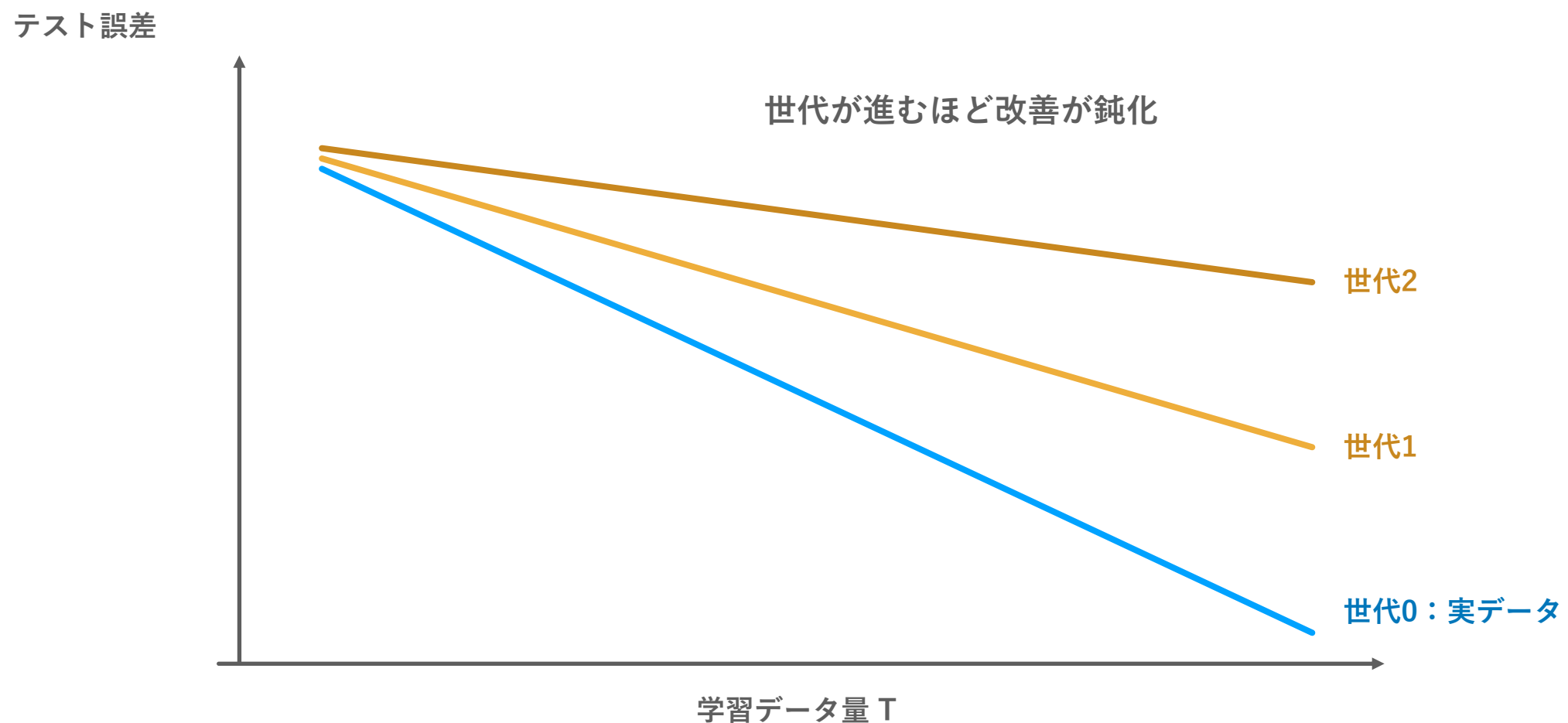
- 合成データのみ：誤差が頭打ち + 獲得済みスキルの「忘却」
- 実データとの混合：Grokking の出現を確認



実験結果②：Llama2-7B

世代ごとにスケールリング則が劣化

- 2世代目では合成データの有用性が大きく低下
- 実データをわずか2%混ぜるだけで劣化を大幅に緩和
- 理論の予測と同じ Grokking を確認



結論

スケーリング則の変化

- 合成データのテール消失が誤差を頭打ちにする
- 世代を重ねるごとに劣化が蓄積する
- 実データの少量混合が Grokking を起こし崩壊を緩和する
- 実データの価値向上