# 合成データと能動学習を用いた大規模言語モデルへの効率的な知識定着

角谷あおい [1]* 河越淳 [2]*
[1] 株式会社 エル・ティー・エス [2] 株式会社 日立システムズ
aoi.kadoya@lt-s.jp
jun.kawagoshi.ec@hitachi-systems.com

## 概要

近年，大規模言語モデル（LLM）は膨大な Web データを活用することで性能を大幅に向上させている．しかし，実務適用においては，ドメイン固有の知識をモデル内部へ定着させる必要があり，これに対して多様な手法が提案されてきた．また，単一の事実をモデルに確実に学習させるためには，多角的な文脈や表現を 100〜1,000 回程度提示する必要があることが報告されているが，そのようなドメインデータを十分に収集・整理するには多大なコストや制約が伴う．本研究では，LLM に対する新たな知識定着を実現する継続事前学習に対して，複数の合成データ生成手法を適用し，効率的な知識獲得方法を検証した．さらに，能動学習を用いたデータセット削減手法を組み合わせることで，知識定着をより効率化する方法を検討した．実験の結果，事実ベースの合成データ生成手法によって多様な合成データを準備することで，単純な言い換え手法と比較して学習回数を 69 %削減しつつ，モデル内部への知識定着が可能であることが明らかとなった．一方で，能動学習手法による学習効率化は，合成データを用いる環境下では期待した性能改善を示さなかった．

## 1 はじめに

大規模言語モデル（LLM）は近年急速な発展を遂げ，多様なアーキテクチャーや学習戦略を採用した数多くのモデルが開発されている [1]．中でも，OpenAI o1[2] は，物理・化学・生物の分野で博士レベルに匹敵する認識能力を備えているとも報告されており，その高度な推論能力や汎用性が注目されている．一方で，LLM は膨大かつ汎用的なテキストデータを用いて学習されるため，学習時点までに存在する知識には精通しているものの，その後に発生した最新の事象や，特定の領域・産業固有の知識に対応することは困難である．

このような新たな知識や領域固有の情報を LLM に付与するための手段としては，プロンプトチューニング（Prompt Tuning）や RAG（Retrieval-Augmented Generation）[3] が提案されているが，Prompt Tuning では入力が長いほど計算コストが上昇する傾向がある．また，RAG は比較的実装が容易であり，外部データベースを用いることで最新の情報に継続的にアクセスできるため，実務面での活用が期待されている．一方で，あくまで外部知識ベースを参照するにとどまり，モデルパラメーター自体を更新しないため，深い専門知識を内在化した高度な回答には限界があると指摘されている [4]．

これに対して，ファインチューニング（Fine-tuning）はモデルパラメーターの更新を伴う知識の埋め込みを実現し得る手法として期待されている．しかし，単純な Supervised Fine-Tuning（SFT）を用いて新たな知識を効率的に獲得することは難しいことが指摘されている [5]．一方，継続事前学習（Continual Pre-Training）[6] は，既存の LLM を再利用しつつ新たなドメイン知識を追加学習する手法であり，一からモデルを学習する場合（フルスクラッチ開発）と比較して，計算資源や学習時間などのコストを抑えながら知識獲得を実現できる点が注目されている．実際，この手法を用いて開発された特化型モデルも提案されており，高度なドメイン推論や専門的な応答生成など，従来手法では難しかった成果を実現していることが報告されている [7]．

一方，学習に用いるデータの質と量がその効果を大きく左右することも指摘されている [8]．例えば，単一の事実を確実にモデルに学習させるには，多様な文脈や表現を通じて同一事実を提示し，約 100〜

* この 2 人の著者は本研究に等しく貢献した．
These two authors contributed equally to this work.

1,000 回にわたる多角的な露出が必要である [8]．しかし，実務で継続事前学習を適用する際，特定の知識を十分にカバーする文書が十分な量で確保できるとは限らず，また，それらの知識が知識定着に十分な表現形式で存在していることも稀である．

さらに近年，大規模コーパスの確保そのものが困難になる可能性が議論されており [9]，新たなドメインや最新情報に対応した学習データを十分にカバーできない「データ枯渇問題」が懸念されている．このような状況を踏まえ，今後の LLM 活用では，不足するドメインデータを補完し，新たな知識を効率よく学習させる手段として，合成データを活用するアプローチの重要性が一層高まると考えられる．

加えて，膨大な学習データに対して能動学習手法を用いることで，学習コストを抑えつつ必要な知識を効果的に獲得できることが報告されている [10, 11, 12]．

このような背景を踏まえ，本研究では，知識定着の効率的な達成手段として，実際のデータを模倣した「合成データ」の活用と，学習に有用なサンプルを能動的に選択する「能動学習」に着目する．合成データを用いた継続事前学習による知識定着の有効性は示唆されているものの [13]，知識定着という観点で合成データと能動学習を体系的に比較・検証した研究は見当たらない．

そこで本研究では，以下のリサーチクエスチョンに焦点を当て，合成データおよび能動学習がどの程度知識定着の効率性に寄与するのかを明らかにする．

(1) ***合成データの作成方法は，知識定着に必要な学習回数に，どの程度影響を与えるか？***

(2) ***能動学習によるデータ選抜は，合成データにおける知識定着に必要な学習回数を削減可能か？***

# 2 手法

## 2.1 合成データ生成

LLM の新たな知識定着を図るために，これまで数多くの合成データ生成手法が提案されている．例えば，単なる言い換え [14]，単語ベース [15]，進化的手法 [16]，事実ベース [15]，ペルソナベース [17] など多岐にわたるが，なかでも，グラフベース手法 [13] はドメイン内の概念間関係を構造化して抽象的に再現することにより，モデルが多面的に情報を学習できる点に重点を置いている．これらの手法は主に SFT などの文脈で用いられ，モデルの性能向上や知識定着における有効性が検証されている [14, 15, 16, 17]．また，単語レベルでの拡張よりも事実ベースの合成データを用いたほうが知識定着に有利であることも指摘されている．

本研究では，上記の複数手法のうち，単なる言い換え，事実ベース，およびグラフベースの 3 種類を選定した．これらは，LLM への効率的な知識定着を目的とする本研究の方針に基づき選定したものである．特に「単なる言い換え」をベースライン手法として位置付けることで，他の合成データ生成手法との比較を行う．

- 単純な言い換え [14]
  元の文章を「Easy」「Medium」「Hard」「QA」の 4 種類の難易度・形式に基づいてパラフレーズし，各形式に応じて内容の抽象度や回答形式を変化させる手法．
- 事実ベースの言い換え [15]
  元文章から文中で核となる情報を抽出し，それらをもとに新たな文章を構築する手法．
- グラフベースの言い換え [13]
  元文章からエンティティを抽出し，異なる視点からエンティティ間の関係性を再編成して文章化する手法．

## 2.2 能動学習

データを削減するための先行研究としては，モデル自身の出力を用いてデータ選抜を行う手法 [10, 11] や，データセットの多様性を指標として活用する手法 [12] が報告されている．これらの手法は，学習データの冗長性を排除しながら，必要な情報を効率的に獲得することをめざしており，合成データや能動学習手法を組み合わせることで，さらに効率的な知識定着を実現できると考える．

本研究では以下の 2 つの手法に関して，合成データ環境下におけるデータ選抜効果を検証した．

- 不確実性サンプリング（Perplexity）[10, 11]：
  LLM の損失に対する指標となる Perplexity の程度に応じてデータを選抜することで，学習データ量を削減しつつモデルの精度を向上させる手法．
- 多様性サンプリング（D4）[12]：
  k-means クラスタリング後に，クラスター内の意味的類似度が高いペアデータを削減する

SemDeDup[18] と，k-means クラスタリング後に，セントロイドに近い代表的なプロトタイプデータを削除する SSL Prototypes[19] を順に実行することで，多様性を持つようにデータを選抜する手法.

## 2.3 データセット

本研究では，産業での利用場面を想定し，利用規約・製品情報・作業手順・社内用語・企業情報の 5 カテゴリーのデータを取り扱う．これらのカテゴリーは既存研究 [20] を参考にテンプレート化し，その一部を穴あき形式とすることで GPT-4o が企業の経常利益や製品名，特徴などを自動補完できるようにしている．ただし，GPT-4o を直接利用すると既存企業名や実在製品名が挿入される可能性があるため，生成後に固有名詞や特徴を架空のものへと置き換える工程を追加している．

なお，本研究で生成した文書は平均約 125 トークンで，合計 50 件を作成した．また，これらの文書はそれぞれ，生成された情報をもとに 8 種類の質問（例：経常利益、製品情報、作業手順など）に答えられるよう統一した形式で作成している．

## 2.4 実験設定

本研究の実験では，既存の日本語モデルである elyza/Llama-3-ELYZA-JP-8B[21] をベースモデルとして採用し，学習回数の増加に伴う知識定着度合いを評価する．具体的には，まず節 2.3 で構築した文書データセットを種データとし，calm3-22b-chat[22] を用いて合成データを生成した．プロンプト設計は各先行研究で提案されているものを踏襲し，元文書 1 件あたり 1,000 件の合成データを作成している．

モデルの学習が完了した段階で，5 つのカテゴリーに属する質問（合計 400 件）に対するモデルの応答を GPT-4o により評価する．これにより，モデルが新たな知識をどの程度獲得・保持しているかを定量的に測定する．なお，回答生成時の temperature は予測の安定性を考慮して 0.1 に固定した．

学習時のハイパーパラメーターとしては，batch_size を 4 とし，1 ステップあたり 50 件の合成データを 1 回ずつ処理する形を採用した．学習率は 1.0e-5 とし，cosine スケジューラーを用いて学習ステップが 10 万に達するまでに学習率が 0 になるよう設定している．これらの条件下で継続事前学習（Continual Pre-Training）を行い，各学習段階における知識定着度合いを測定した.

# 3 結果

## 3.1 合成データ生成手法による学習回数とモデル正答率の比較

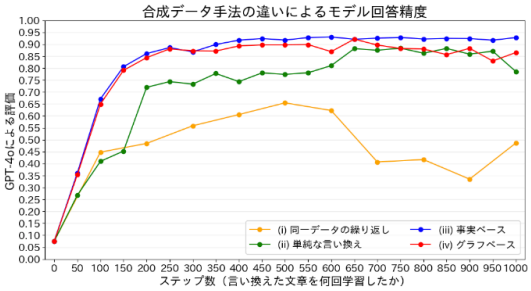


**図 1** 合成データ手法の違いによるモデル正答率

図 1 は，(i) 合成データを用いずに学習を行った場合と，(ii)〜(iv) 各種合成データ生成手法を適用した場合の学習回数に対する精度推移を示している．実験条件 (i) においては，約 500 回の学習を経て最大で約 65 %の精度に達したものの，その後は目立った向上が見られず，長期的に精度が 30 %程度まで低下した．これにより，限られたデータを単純に反復利用しただけでは，知識定着には限界があり，むしろ過学習により精度低下を招く可能性が示された.

一方，合成データを活用する手法 (ii)〜(iv) については，学習回数を増やすほど精度が着実に向上し，最終的には 90 %程度の精度を達成した．評価精度 85 %に到達するまでの学習回数を比較すると，(ii) の単純な言い換え手法では約 650 回を要したのに対し，(iii) のグラフベース手法は約 250 回，(iv) の事実ベース手法は約 200 回で到達した．さらに，最終的に 1,000 回の学習を行った際の精度を比較すると，(ii) の単純な言い換え手法は約 90 %未満で頭打ちとなったのに対し，(iii) のグラフベース手法は 92 %程度まで向上し，(iv) の事実ベース手法は 93 %を超える精度を安定的に維持している．以上の結果から，合成データの活用は，知識定着に効果的であるのに加えて，学習効率にも効果があることが確認された.

## 3.2 能動学習手法による学習回数とモデル正答率の比較

能動学習（Perplexity，D4）を用いた学習データの削減による知識定着への実験結果を図 2 に示す.

Perplexity によるデータ削減は，全体の上位 50%（各文章に対して生成した 1,000 件の合成データの内，Perplexity が上位 500 件のデータ）を選択した場

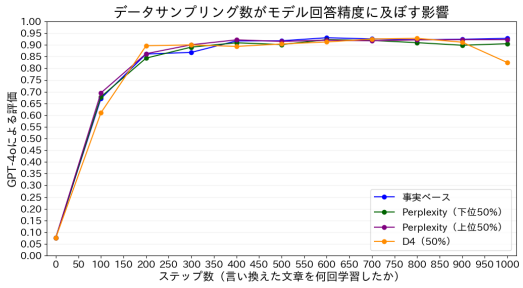


**図 2**　Perplexity，D4 によるサンプリング削減効果

合，下位 50%を選択した場合，どちらも知識定着までの学習回数に大きな影響を与えないことが示された．

D4 によるデータ削減に関しても，図 2 に示されるように，Perplexity によるサンプリングと同様，知識定着までの学習回数に大きな影響を与えないことが示された．

## 4　考察

***合成データの作成方法は，知識定着に必要な学習回数に，どの程度影響を与えるか？***

単一データの反復では十分に知識定着ができなかったことから，限られたデータを同一形式で繰り返し使用する手法では，過学習が生じやすく，高い精度を達成することは難しいと考えられる．

一方，合成データを用いた手法はいずれも，学習を重ねるほど精度を着実に高められることが明らかとなった．このことは，データの多様性や情報量を人工的に補完することで，モデルがより幅広い文脈・表現を学習できるようになるためと考えられる．特に，事実ベース手法は総じて最高水準の精度を達成し，学習回数という観点からも効率的に性能を向上させられることが確認された．一方，答えが一意に定まる詳細情報（製品情報カテゴリー）に関しては，グラフベース手法が最良の性能を示している（Appendix A 参照）．これは，製品情報のように階層的かつ一意に定義される知識領域では，グラフ構造を活用することで情報を体系的に表現しやすい一方，多様な文脈を含むより包括的な領域では，事実ベースのほうが必要とされる情報を包括的に提供できる可能性があることを示唆する．

以上の考察から，合成データ生成手法は特定の知識ドメインに合わせて適切なものを選択する必要があり，汎用的に最も優位となる手法が存在するとは限らないことが示唆される．今後は各ドメインやタスクにおけるデータ分布・構造をより詳細に分析し，最適な合成手法やその組み合わせを見極めることが，継続事前学習のさらなる効率化および精度向上において重要な課題になると推測する．

***能動学習によるデータ選抜は，合成データにおける知識定着に必要な学習回数を削減可能か？***

Perplexity や D4 といった能動学習手法は，先行研究においては一定の効果を示すことが報告されている一方で，合成データに対しては効果が限定的であることが明らかになった．

その理由として，一般的に能動学習の効果が確認されている Web データは，品質の高いデータと低いデータが混在しているのに対し，合成データは品質の高いデータが多く含まれるため，学習効果の高いデータのみを選抜する手法の効果が十分に発揮できないと考えられる．また，合成データは似通ったデータが生成される傾向があるため，データ選抜による効果が限定的となった可能性がある．

このような背景から，1 つの LLM に対して多様な合成データを生成させる場合には，データ選抜を重視するよりも，1 つの事実に対し，200 件以上のデータを生成し（Appendix B 参照），それらのデータをすべて活用して繰り返し学習を行うアプローチが効果的であると言える．

## 5　おわりに

本研究では，合成データと能動学習を組み合わせた継続事前学習による新たな知識定着の可能性を探った．その結果，合成データは学習回数の増加とともに学習効率が向上し，特に事実ベース手法が安定して高い効率を達成した．一方，能動学習手法は，すでに品質が高い合成データ環境下では限定的な効果しか示さなかった．これらの知見は，ドメイン知識の学習が求められる産業利用において，合成データを設計・活用する際の指針として有用である．今後は，より多様なドメインやタスクを対象に合成データと能動学習の適用範囲を検証し，データ特性やモデル特性に応じた最適な学習戦略を構築することが重要だと考えられる．

## 謝辞

本研究は，株式会社日立システムズと株式会社エル・ティー・エスの共同研究により実施した．また，本研究の実験環境には，株式会社フィックスターズの Fixstars K4 クラウドを活用しました．

# 参考文献


[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2024.

[2] OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.

[4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. **arXiv preprint arXiv:2312.10997**, 2023.

[5] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? **arXiv preprint arXiv:2405.05904**, 2024.

[6] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. **arXiv preprint arXiv:2302.03241**, 2023.

[7] Masanori Hirano and Kentaro Imajo. The construction of instruction-tuned llms for finance without instruction data using continual pretraining and model merging. **arXiv preprint arXiv:2409.19854**, 2024.

[8] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. **arXiv preprint arXiv:2404.05405**, 2024.

[9] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. **arXiv preprint arXiv:2211.04325**, pp. 13–29, 2024.

[10] Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale, 2023.

[11] Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L. Leavitt, and Mansheej Paul. Perplexed by perplexity: Perplexity-based data pruning with small reference models, 2024.

[12] Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. D4: Improving llm pretraining via document de-duplication and diversification, 2023.

[13] Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining, 2024.

[14] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling, 2024.

[15] Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. Injecting new knowledge into large language models via supervised fine-tuning, 2024.

[16] Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. Automatic instruction evolving for large language models. **arXiv preprint arXiv:2406.00770**, 2024.

[17] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. **arXiv preprint arXiv:2406.20094**, 2024.

[18] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

[19] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023.

[20] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction, 2024.

[21] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[22] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.

## A　各合成手法におけるカテゴリーごとの評価スコア

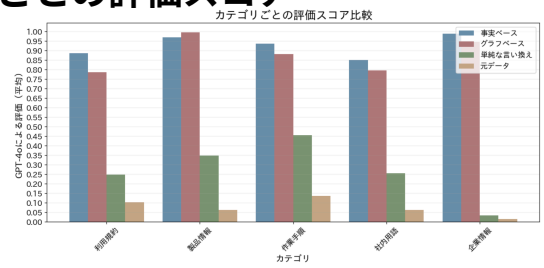


**図 3**　カテゴリーごとの評価スコア比較

各合成手法においてカテゴリーごとに分析した結果，総じて事実ベース手法はグラフベース手法を含む他の合成データ生成手法を上回る性能を示したものの，製品情報カテゴリーに限ってはグラフベース手法の方が高精度を達成していることが確認された．製品情報カテゴリーでは，

> {product} は，{year} 年に発売された製品で，{charactor} が特徴です．この製品は主に，{place} で，{use} 利用します．〜”

といった形式で，{product}，{year}，{charactor}，{place}，{use}，{previous_product}，{times}，{power_consumption}，{weight} などの変数を用いて，架空の製品情報を細部まで定義している．

このように階層的かつ一意に定義可能な属性が多い場合，グラフ構造を用いた手法が情報の関連性を体系的に捉えやすく，高い精度を発揮する要因の一つと考えられる．

## B　合成データの選抜数による影響

1 文章に対して生成した 1,000 件の合成データからサンプリングするデータ数を変化させた場合の，学習回数と回答精度の結果を図 4 に，学習時における訓練誤差を図 5 に示す．

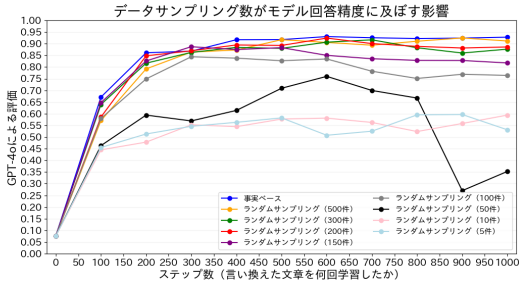


**図 4**　合成データサンプリング数による精度の変化

図 4 に示すように，1 つの文章に対して 200 件以上合成データをサンプリングした場合，600 ステップ以内に 90% 以上の精度を達成する．一方で，合成データ数が 150 件以下の場合は，1,000 ステップの学習を実施しても 90%以上の精度を達成できないことが確認された．

この結果は，図 5 に示すように，1 つの文章（事実）を言い換えたデータの数が少ない場合は，訓練誤差が 0 に近づき，過学習が発生することが原因であると考えられる．

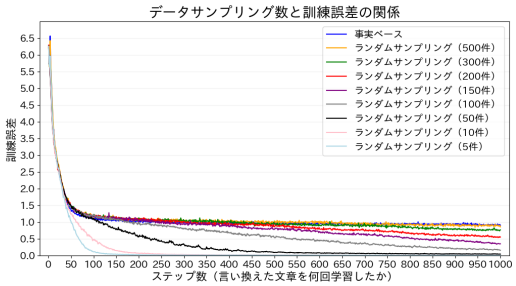


**図 5**　合成データサンプリング数による訓練誤差の変化

## C　少数データの選抜効果

図 6 に，1 文章に対して生成した 1,000 件の合成データから，100 件をサンプリングした場合の削減効果を示す．

サンプリング数が 100 件と少数の場合，Perplexity が低いデータをサンプリングすることで精度が向上する一方，Perplexity が高いデータや多様性を重視したデータをサンプリングすると精度が低下する傾向が見られた．

サンプリングされたデータを調査した結果，Perplexity が高いデータには，英語のみで記述された不自然なデータが数件含まれていた．また，多様性サンプリングを用いた場合も，多様性を保つために同様の不自然な文章が含まれており，このような質の低いデータが含まれることが精度低下の原因であると考えられる．

一方，図 7 に示すように，1 文章あたり 200 件のデータをサンプリングした場合，サンプリング手法による精度への影響は小さい．これは，LLM が生成した合成データの大部分が質の高いデータであり，どの手法を用いても質の高いデータが大多数を占めるため，手法による違いが精度に与える影響が小さくなったと考えられる．

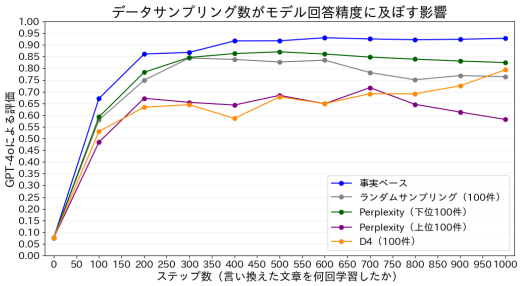


**図 6**　1 文章あたり 100 件のデータを選抜した場合の影響

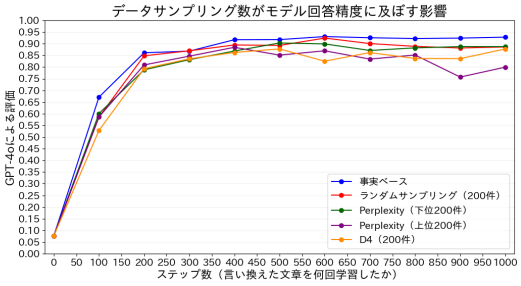


**図 7**　1 文章あたり 200 件のデータを選抜した場合の影響