双方向推論とデータ拡張を両立する空白文字順序復元型 LLM 学習フレームワーク

高　鵬挙 [1]　山崎　智弘 [1] 岩田　憲治 [1]
[1] 株式会社東芝　総合研究所
AI デジタル R&D センター　コラボレイティブ AI 研究部
{pengju.gao.t15,tomohiro.yamasaki.n90,kenji.iwata.h13}@mail.toshiba

概要

本研究は、自己回帰型 LLM の「未来文脈を活用できない」という構造的制約と、専門ドメインにおけるデータ不足問題を同時に解決するため、新しい学習フレームワークを提案する。過去と未来の文脈を同時に参照する必要がある、並べ替えた文字を復元するタスクを自動生成し、対象範囲を動的に変えることで大量の学習データを生成する。さらに、文字順序復元タスクと、文の前半を与えて後半を生成させる文章継続タスクを混合し、混合比率 m を変えたときの性能変化を定量化した。金融・医療ベンチマークにおいて、zero-shot で平均+1.0 ポイント、two-shot 医療で+3.2 ポイントの改善を達成した。

1　はじめに

近年、GPT-3 [1] や LLaMA 系列 [2] に代表される自己回帰型の大規模言語モデル（LLM）は、次トークン予測に基づく単方向学習により高い生成性能と指示追従能力を獲得してきた。しかし、この構造には制約がある。単方向学習では未来の文脈を過去の推論に利用できず、文全体の整合性や逆方向の理解が必要な場面で性能が低下する。例えば、契約文の欠損補完、医療記録の順序整合、トラブル事例の原因・結果分析など、双方向の文脈参照が不可欠なタスクにおいて、現行モデルは信頼性に課題を抱えている。

さらに、金融や医療といった専門ドメインではラベル付きデータの不足が深刻であり、追加データの整備には多大なコストが必要となる。このため、双方向推論を強化しつつ、データ不足を緩和する手法が求められている。

本研究では、汎用テキストから自動生成可能な「文字順序復元タスク」を提案する。このタスクは、ランダムに挿入した空白と文字乱序ヒントを用いて、過去と未来の文脈を同時に参照する推論を促す。また、空白位置や長さを動的に変化させることで、大量の疑似ラベルを生成し、専門ドメインにおけるデータ不足を補う。さらに、文字順序復元タスクと文章継続タスクを混合し、混合比率 m を変化させた場合の性能を定量的に評価することで、双方向推論信号の寄与を分析する。

2　関連研究

自己回帰型言語モデル（GPT-3[1], LLaMA 系列[2]）は因果構造に基づく次トークン予測を採用し、高い生成性能を示す一方、未来文脈を利用できないため双方向推論が困難である。これに対し、マスク型モデル（BERT[3], T5[4]）は双方向文脈を扱えるものの、生成には別途デコーダを必要とし、推論速度や汎用性に制約がある。

XLNet[5] や MPNet[6] は順序ロバスト性を強化したが計算コストが高い。Fill-in-the-Middle(FIM)[7] や BART[8] は文中間補完やノイズ除去で双方向性を付与するが、データ生成コストが課題である。

順序判定や復元に焦点を当てた手法として、Shuffled-token detection[9] はトークン順序の正誤判定を学習し、TOR[10] は順序回復を明示的に目的化する。SPT[11] や TOP[12] は部分的な順序制約を課し、MTP[13] はマルチターゲット復元でロバスト性を向上させている。

本研究はこれらの着想を踏まえつつ、ランダム空白挿入＋文字乱序ヒント付き復元タスクを自動生成し、双方向推論信号を付与する点、さらに混合比率 m を段階的に変化させて性能寄与を定量化する点で新規性を有する。

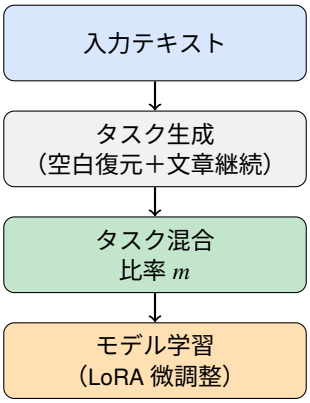


図 1　提案手法の全体フレームワーク

表 1　指示データ例（Instruction/Output は一部抜粋）

金融	
ドメイン	**金融**
タスク種別	空白復元
サンプル内容	**Instruction:** ヒント 空欄 1（5 文字）=者, 保, 約, 契, 険／本文「金融庁は預金者、__、有価証券の投資者等の保護…」。 **Output:** 空欄 1=保険契約者
医学	
ドメイン	**医学**
タスク種別	文章継続
サンプル内容	**Instruction:** 身体所見（身長 165cm、BMI16.5、胸部所見正常など）の冒頭に続く 100 文字以内の自然な継続を作成。 **Output:** 呼吸音・心音に異常なく、体幹と四肢に近位筋優位の筋萎縮を認める…

3　提案手法

本節では，双方向推論信号とデータ拡張を同時に実現する学習フレームワークを説明する。対象は日本語の汎用テキストおよび専門ドメインテキストである。本手法は，空白復元タスクと文章継続タスクを混合し，混合比率 m を調整することで双方向推論の寄与を分析する。

3.1　全体フレームワーク

図 1 に示すように、本手法は次の 4 ステップで構成される：

1. 入力テキストを取得。
2. 空白復元タスクと文章継続タスクを生成。
3. 混合比率 m に従ってタスクを統合。
4. LoRA によるモデル微調整を実施。

この流れにより、双方向推論信号を付与し、動的な空白設定で大量の疑似ラベルを生成し、専門ドメインのデータ不足を緩和する。

3.2　空白復元タスク

空白復元タスクでは，文中の複数箇所を下線に置換し，それぞれの語を文字単位で乱序化したヒントと文字数を提示して原文を再構成させる。乱序ヒントにより，単なる穴埋めではなく，順序推定能力を強制的に活性化する。

具体的には，形態素解析で語スパンを抽出し，句読点・記号を除外した候補から複数箇所を選択する。各空白は最小長を満たし，空白間には一定の距離を確保して局所的な欠損集中を防ぐ。選択語句は表層文字をランダム並べ替えたヒントとして提示し，回答は「空欄番号＝語句」を縦棒で連結する厳密書式のみを許可する。

3.3　文章継続タスク

文章継続では冒頭数文を条件として与え，100 文字以内で自然な続きのみを生成させる。これにより，文脈を保持した自然な展開を学習する。$m = 0$ の場合，文章継続のみで追加学習するの意味である。

3.4　学習の混合

1 エポック内で空白復元と文章継続の 2 タスクを混合比率 m でサンプリングし、損失は以下で定義する：

$$\mathcal{L}_{\text{total}} = m\,\ell_{\text{blank}} + (1-m)\,\ell_{\text{cont}}.$$

ここで、ℓ_{blank} は空白復元タスクの平均損失、ℓ_{cont} は文章継続タスクの平均損失を表す。

4　実験設定

本研究は金融と医学二つドメインで空白復元と文章継続の混合比率 $m \in [0, 1]$ を 0.1 刻みで変更し，ベースモデルを訓練した。$m = 0$ は文章継続のみ，$m = 1$ は空白復元のみの学習を意味する。訓練できたモデルの評価は二つドメインのベンチマークで行った。

4.1　モデル訓練

以下のコーパスを用いて，モデルを訓練した。

- **金融ドメイン**：自社で収集したコーパス（約 67 万行のテキスト）。
- **医学ドメイン**：公開症例報告 iCorpus[14] を利用（179 件の症例報告）。

表 1 に両ドメインから抽出したサンプルを示す。サンプルは各ドメインの代表的な設問からランダ

ム抽出し，難易度と語彙分布の偏りを避けるよう調整した。ヒント列と指示テンプレートは一定の体裁で整形されており、そのまま LoRA 微調整用の指示データとして利用可能である。学習条件を表 2 に示す。

4.2 評価ベンチマーク

金融： 日本語金融ベンチマーク [15] に基づき，cma_basics, fp2, security_sales_1, cpa_audit, chabsa の 5 タスクを選定。これらは金融知識を要する多肢選択問題で構成され，長文・契約文・計算問題などに対して空白復元による文脈統合が有効か検証する。

医学： 日本語バイオメディカル LLM 評価用ベンチマーク JMedBench[16] を採用し，medmcqa, usmleqa, medqa, mmlu_medical, jmmlu_medical, igakuqa, pubmedqa の 7 タスクを評価。臨床推論や専門用語理解を含む設問を中心に構成し，選択式・短答式の正答率（Accuracy）で比較し，平均値を「ドメイン平均」として報告。

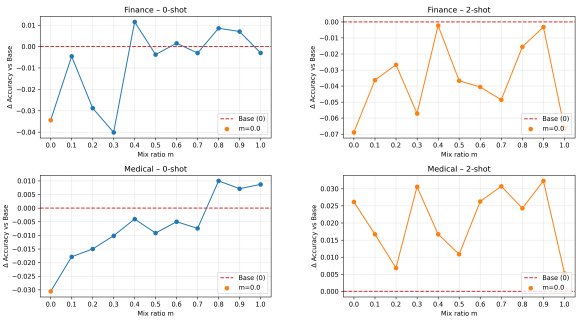
各タスクは zero-shot（指示のみ）と two-shot（サンプル 2 件提示）の 2 設定で推論。two-shot は few-shot 効果検証のため採用（1-shot 不足・3-shot 高コストを考慮）。

表 2 学習条件の概要

項目	設定
Base	Llama-3-ELYZA-JP-8B
LoRA 適用層	鍵・問い・値＋最終線形層
LoRA 設定	ランク 16，スケール 16
学習率	2×10^{-4}
最大シーケンス長	1,024
スケジュール	コサイン＋ウォームアップ
バッチサイズ	バッチサイズ 4（累積で 32））

4.3 比較対象

- **Base**：公開済み LoRA 事前調整モデル（`base.json`）。
- $m = 0$：文章継続のみで追加学習した消融モデル。
- $m = 0.1$〜1.0：空白復元と文章継続を混合した提案モデル。タスク平均が最も高い設定を主結果として報告。



(a) 0-shot 設定　　(b) 2-shot 設定

図 2 混合比率 m と性能変化（Base 比）の関係。図 (a) は 0-shot、図 (b) は 2-shot 設定を示す。横軸は混合比率 m（0.0〜1.0）、縦軸は Base モデルに対する正答率差分（Δ Accuracy）。

5 実験結果

金融・医療ドメインにおける QA タスクで混合比率 m の影響を評価した。評価は zero-shot および two-shot 設定で正答率（Accuracy）を比較し、m は 0.0〜1.0 を 0.1 刻みで調整した。

性能は混合比率に対して非線形に変化し、金融では $m \approx 0.4$、医療では $m \approx 0.8$〜0.9 で最大改善を示した（図 2）。特に医療ドメインでは two-shot 設定で全体的に Base モデルを上回り、双方向推論タスクの寄与が明確である。

表 3 に金融ドメインの 5 タスク別の正答率と平均値を示す。zero-shot 設定では，Base モデルの平均正答率は 0.406 であるのに対し，文章継続のみで追加学習した $m = 0.0$ は 0.371 と-3.5 ポイント低下した。一方，空白復元を 40%混合した $m = 0.4$ では平均 0.417 となり，Base 比+1.1 ポイント，$m = 0.0$ 比+4.6 ポイント改善した。特に長文化契約文を含む `security_sales_1` では，Base 比+10.5 ポイント（$m = 0.0$ 比+5.2 ポイント）と顕著な改善を示した。一方，既存知識依存度が高い `cma_basics` や `chabsa` では差分が小さい。

two-shot 設定では，Base（0.434）が最良だが，提案モデル $m = 0.4$ も 0.432 でほぼ同等である。$m = 0.0$ は 0.365 で Base 比-6.9 ポイントと大きく低下した。なお `security_sales_1` では $m = 0.4$ が Base 比+8.8 ポイントと改善している。

表 4 に医療ドメインの 7 タスク別の正答率と平均値を示す。zero-shot 設定では，Base モデルの平均正答率は 0.442 であるのに対し，$m = 0.0$ は 0.412 と-3.0 ポイント低下した。一方，空白復元を 80%混合した $m = 0.8$ では平均 0.452 となり，Base 比+1.0

表 3　金融タスク別＋平均正答率（0-shot と 2-shot）

	0-shot			2-shot		
タスク	Base	m=0.0	m=0.4	Base	m=0.0	m=0.4
cma_basics	0.342	0.237	0.316	0.421	0.211	0.316
fp2	0.248	0.257	0.253	0.248	0.183	0.259
security_sales_1	0.351	0.404	0.456	0.404	0.404	0.491
cpa_audit	0.173	0.163	0.176	0.173	0.188	0.181
chabsa	0.914	0.796	0.886	0.924	0.840	0.912
平均	0.406	0.371	0.417	0.434	0.365	0.432
Δ vs Base	—	-0.035	+0.011	—	-0.069	-0.002

表 4　医療タスク別＋平均正答率（0-shot と 2-shot）

	0-shot			2-shot		
タスク	Base	m=0.0	m=0.8	Base	m=0.0	m=0.9
medmcqa	0.364	0.329	0.370	0.235	0.328	0.305
usmleqa	0.375	0.352	0.386	0.331	0.268	0.334
medqa	0.320	0.303	0.332	0.278	0.267	0.284
mmlu_medical	0.481	0.445	0.476	0.463	0.436	0.454
jmmlu_medical	0.477	0.468	0.494	0.455	0.435	0.447
igakuqa	0.433	0.389	0.419	0.331	0.368	0.355
pubmedqa	0.647	0.597	0.690	0.488	0.662	0.628
平均	0.442	0.412	0.452	0.369	0.395	0.401
Δ vs Base	—	-0.030	+0.010	—	+0.026	+0.032

ポイント改善した。pubmedqa では Base 比+4.3 ポイントと顕著な伸びを示したが，知識記憶依存度が高い igakuqa や mmlu_medical ではわずかな低下が見られる。

two-shot 設定では，$m = 0.0$ でも平均 0.395 で Base（0.369）を+2.6 ポイント上回るが，$m = 0.9$ では 0.401 となり Base 比+3.2 ポイント，$m = 0.0$ 比+0.6 ポイントの追加改善が確認された。推論・読解寄りの課題で伸長が見られる一方，medmcqa, igakuqa, pubmedqa では若干低下している。

6　考察

図 2 から、性能は混合比率 m に対して非線形に変化することが分かる。金融ドメインでは $m \approx 0.4$ 付近、医療ドメインでは $m \approx 0.8$〜0.9 で最大改善を示し、タスク特性に応じた最適比率の存在が確認された。

空白復元タスクは、単なる穴埋めではなく乱序ヒントを伴うことで順序推定能力を強制的に活性化し、双方向推論を促進する。特に長文整合や欠損補全を含む設問で顕著な効果があり、金融では security_sales_1 で最大+10.5 ポイント、医療では pubmedqa で+4.3 ポイント改善した。これは、未来文脈を利用する学習機会を提供し、因果構造の片方向性を補完することを示している。

一方、$m = 0.0$（文章継続のみ）は双方向信号が欠落し、長距離依存や選択肢整合を要する設問で精度が低下した。文章継続タスクは「過去→未来」への単調な展開に偏り、逆方向の整合性を学習できないためである。また、損失分布が局所的生成に集中し、長文全体の整合性を促す信号が不足することも一因と考えられる。

金融では中間比率で安定した改善が見られ、契約文や反転質問など文脈統合が必要なタスクで効果が大きい。一方、医療では高比率で改善が最大となり、few-shot との補完効果が強く、推論・読解寄りの課題で伸長が確認された。知識記憶依存度が高いタスク（cma_basics, igakuqa など）では改善が限定的であり、双方向推論信号よりも知識量が支配的であることが示唆される。

7　おわりに

本研究では，ランダム空白挿入と文字乱序ヒントを用いた自己教師タスクを軸に，双方向推論強化とデータ拡張を両立する学習フレームワークを提案した。空白復元タスクと文章継続タスクを混合し，混合比率 m を調整することで，金融・医療ドメインにおいて zero-shot 平均+1.0 ポイント，two-shot 医療+3.2 ポイントの改善を確認した。これにより，因果構造に起因する片方向性の制約を緩和し，専門ドメインにおけるデータ不足を効果的に補完できることを示した。

今後の課題として，以下を検討する：

- **混合比率の自動最適化**：タスク特性に応じて m を動的に調整するアルゴリズムの導入。
- **双方向性強化タスクの追加**：Fill-in-the-Middle（FIM）や順序不変正則化を組み込み，文脈統合をさらに強化。
- **カリキュラム学習の適用**：空白スパンの長さ・数を段階的に難化させ，学習の安定性と性能向上を図る。

本フレームワークは，専門ドメインにおけるデータ不足問題を緩和しつつ，双方向推論能力を高める有効なアプローチである。今後は、より多様なドメインへの適用を進めるとともに、生成品質および推論精度のさらなる向上を目指す。

商標について:

"GPT-3"及び"LLaMA"はそれぞれ OpenAI 社、Meta 社の登録商標です。またその他本稿に掲載の商品、機能等の名称は、それぞれ各社が商標として使用している場合があります。

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.

[2] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of NAACL-HLT**, pp. 4171–4186, 2019.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[5] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In **Advances in Neural Information Processing Systems**, pp. 5754–5764, 2019.

[6] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In **Advances in Neural Information Processing Systems**, 2020.

[7] Mohammad Bavarian, Aakanksha Chowdhery, Yury Lee, Sharan Narayan, Jacob Devlin, Pavel Izmailov, Anselm Levskaya, Alexandru Salcianu, Blake Hechtman, Noam Shazeer, et al. Efficient training of language models to fill in the middle. **arXiv preprint arXiv:2207.14255**, 2022.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of ACL**, pp. 7871–7880, 2020.

[9] Subhadarshi Panda, Uma Jain, Guanghui Qin, and Jiawei Han. Shuffled-token detection for refining pre-trained roberta. In **Proceedings of NAACL Student Research Workshop**, pp. 88–93, 2021.

[10] Ridouane El Mesbahi, Andrei Omelianenko, Wilker Aziz, Ivilin Stoianov, and Ivan Titov. On the utility of enhancing bert syntactic bias with token reordering pretraining. In **Proceedings of CoNLL**, pp. 165–182, 2023.

[11] Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. Mitigating reversal curse in large language models via semantic-aware permutation training. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 11453–11464, 2024.

[12] Ahmed Zuhri, Luyu Zhang, Nicolas Carion, et al. Predicting the order of upcoming tokens improves language modeling. **arXiv preprint arXiv:2508.19228**, 2025.

[13] Timur Aynetdinov and Alan Akbik. Pre-training curriculum for multi-token prediction in language models. **arXiv preprint arXiv:2505.22757**, 2025.

[14] 東京大学大学院医学系研究科医療 AI 開発学講座. 症例報告コーパス（icorpus）, 2021. https://minds.juntendo.ac.jp/icorpus.

[15] Masanori Hirano. Construction of a Japanese Financial Benchmark for Large Language Models. In **Joint Workshop of the 7th Financial Technology and Natural Language Processing (FinNLP), the 5th Knowledge Discovery from Unstructured Data in Financial Services (KDF), and The 4th Workshop on Economics and Natural Language Processing (ECONLP)**, pp. 1–9, 2024.

[16] Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. Jmedbench: A benchmark for evaluating japanese biomedical large language models, 2024.