自己生成・自己選好データを活用した事後学習の効果検証

吉田 希世 [1]　Haocheng Zhu[1]　王 天奇 [1]　鈴木 潤 [1,2,3]
[1] 東北大学　[2] 理化学研究所　[3] 国立情報学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

大規模言語モデルの事後学習に利用するデータは，一般に高品質であることが望ましい．従来は人手で作成したデータが積極的に用いられてきたが，近年は，コスト削減等の動機でモデルが生成したデータを活用して性能向上を図るアプローチも注目を集めている．本研究では，このアプローチの中でも，外部モデルや人手による評価に依存せずに，学習対象の言語モデル自身が生成し，選好した自己生成データのみを用いた事後学習に焦点を当て，性能向上に寄与する手法を模索した．実験の結果，学習対象モデル自身が生成・選好したデータを用いた事後学習が，適切な学習枠組みと組み合わせることで有効に機能する可能性が示唆された．

1　はじめに

事前学習を終えた大規模言語モデルに対して，教師あり微調整 (Supervised Fine-Tuning; SFT)[1] や直接選好最適化 (Direct Preference Optimization; DPO)[2] といった事後学習を施すことにより，モデルの指示追従能力の向上や人間の選好に沿った応答生成が可能となっている．これらの事後学習に用いられるデータは，モデルに適切な振る舞いを学習させるため，高品質であることが望まれ，従来は人手で作成したデータが積極的に用いられてきた．しかし，人手によるアノテーションは高いコストを伴い，スケーラビリティに欠けるという課題がある．この課題に対し，近年では言語モデルが生成したデータを事後学習に活用して性能向上を図るアプローチが注目を集めている．

しかし，既存研究の多くはより高性能な外部モデルが生成したデータを学習対象のモデルが模倣する蒸留のアプローチになっており [3]，性能の上限が外部モデルに依存する可能性がある．また，計算資源やライセンス制約によって，高性能な外部モデルを利用できず，利用可能な単一モデルのみを用いて学習を完結させたいという状況も想定される．このような状況下では，外部モデルに依存せず，学習対象であるモデル自身の能力のみで性能向上を図ることが可能な事後学習手法が求められる．

このような背景から，我々は，学習対象の言語モデル自身が生成したデータを活用する事後学習に着目した．具体的には，学習対象の言語モデルが単一の入力に対して複数の応答候補を生成し，それらを同一モデルによって選好することで，新たな事後学習用データセットを構築する．本研究では，応答の選好基準として最小ベイズリスク (Minimum Bayes Risk; MBR) に基づくスコアを用いる．以降，本稿では MBR スコアと呼ぶ．MBR スコアは複数の候補の中から期待損失を最小化するという原理に基づく指標である．先行研究では，MBR スコアの算出にあたって，参照文に基づくニューラル評価器 (BLEURT[4]) や，意味的類似度指標 (BERTScore[5])，さらに外部モデルを用いた LLM-as-a-Judge[6] による評価スコアが主に用いられている [7, 8]．

これに対して，本研究では，応答生成だけでなく MBR スコアの算出も学習対象のモデル自身で行う．すなわち，本研究の目的は，単一の言語モデルによる自己生成および自己選好データを用いた事後学習によって，外部モデルに依存せずにモデルの性能向上に寄与する方法を模索することである．

2　関連研究

MBR は，複数の候補の中から期待損失を最小化する候補を選択するという原理である．候補集合 $\mathcal{Y} = \{y_1, \ldots, y_N\}$ と損失関数 $L(\cdot, \cdot)$ に対して，MBR による解 y^* は以下のように定式化される:

$$y^* = \arg\min_{y_i \in \mathcal{Y}} \mathbb{E}_{y \sim p(y|x)} \left[L(y_i, y) \right] \tag{1}$$

この原理に基づく MBR デコーディングは従来，機械翻訳を中心に用いられてきており [9, 10]，その有効性が示されている [11, 12, 13]．しかし，MBR デコーディングは単一の最尤解よりも高品質な出力

を生成できる反面，推論時に多数の候補を生成し比較する必要があるためコストが高いという課題がある．この課題に対して，Yang ら [7] は，MBR デコーディングによって選択されたデータを教師データとして用いて再学習する手法を提案した．この手法では，MBR デコーディングにより得られる選好を学習を通じてモデルのパラメータに内在化することで，推論時には通常のデコーディングを用いながらも，MBR デコーディングと同等以上の性能を達成できることが示されている．Wu ら [8] は，MBR デコーディングを指示追従タスクにも適用し，LLM-as-a-Judge や意味的類似度指標を用いた MBR スコアに基づく出力選好が有効であることを示した．さらにその効果を学習によって内在化できることを示した．この研究は MBR に基づく出力選択および学習の枠組みが，機械翻訳に限定されず，より一般的な言語生成タスクに拡張可能であることを示している．

しかしながら，これらの研究において出力の評価には学習対象とは異なるモデルを用いる設定が含まれており，結果として，外部モデルの知識を間接的に蒸留している可能性を完全に排除することはできない．そのため，得られた性能向上が提案手法そのものの有効性に起因するのか，あるいは外部モデルの性能に依存した結果であるのかを切り離して評価することは容易ではない．

これらの先行研究に対して本研究は，MBR に基づく学習の枠組みに着想を得つつも，応答生成と応答選好の双方を単一モデルで完結させる点に特徴がある．外部の高性能モデルや評価器を用いないことで，間接的な知識蒸留の影響を排除し，手法そのものが汎化性能の向上にどの程度寄与するのかを検証する．

3 手法

本研究における学習手順の概要を図 1 に示す．まず，既存のデータセットを用いて事前学習済みモデルに対して事後学習を行い，ベースラインモデルを構築する．次に，このベースラインモデルを用いて既存のデータセットの各データに対し，最後の assistant 応答部分を複数生成する．元の応答 y_i と生成された応答 $\{y_{i1}, \ldots, y_{in}\}$ を合わせたものを応答候補集合 $\mathcal{Y} = \{y_i, y_{i1}, \ldots, y_{in}\}$ とする．その中から，新たな応答を選好し，事後学習用データセットを構築する．

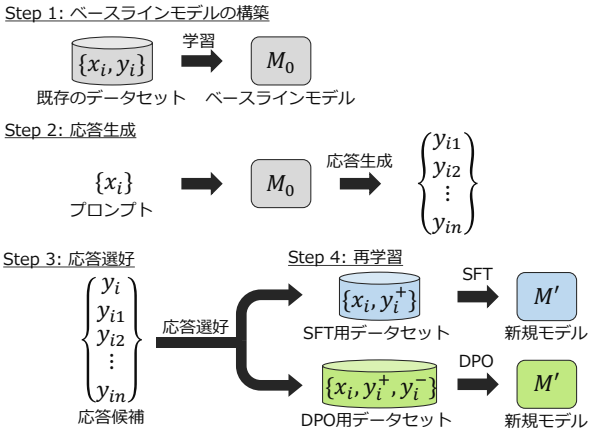

図 1 学習手順の概略図．

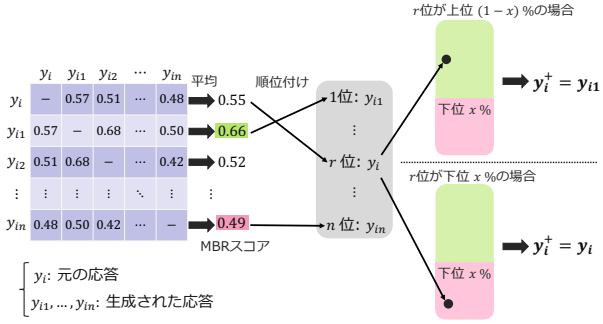

図 2 MBR スコアの算出および閾値による応答選択の概略図．各応答候補間の類似度の平均値を MBR スコアとする．

応答の選好基準には，図 2 に示すように，応答候補間の類似度の平均を取ることで算出される MBR スコアを用いる．類似度指標としては応答候補間の埋め込みベクトルのコサイン類似度を用いる．候補集合 $\mathcal{Y}$ に含まれる任意の候補 y_k の埋め込みベクトルを $\mathbf{h}_k \in \mathbb{R}^d$ とすると，y_k に対する MBR スコアは次式で定義される：

$$\mathrm{MBR}(y_k) = \frac{1}{|\mathcal{Y}| - 1} \sum_{y_l \in \mathcal{Y} \setminus \{y_k\}} \cos(\mathbf{h}_k, \mathbf{h}_l) \qquad (2)$$

したがって，本研究における MBR スコアは，各応答が他の応答候補と平均的にどの程度意味的に近いかを表す指標であり，多数の候補の中で意味的に他の応答と整合的な応答を選択するための基準として機能する．

本研究では，この MBR スコアを元に，以下に示す手法に基づいてそれぞれ新たな事後学習用データセットを構築し，再度学習を行った．

手法 1 (MBR-SFT) プロンプト x_i に対する応答候補集合の中で MBR スコアが最も高いものを新たな正解応答 y_i^+ としてデータセット $\{x_i, y_i^+\}$ を構築し，SFT を行う．ただし，元の応答の MBR スコア

の順位 r が事前に定めた閾値を下回る (候補集合の下位 x %に含まれる) 場合には，元の応答を正解応答として用いる (図 2 参照)．このように，元の応答のスコアが低い場合に正解応答として元の応答を採用する理由は，元のデータがある程度高品質であると仮定した場合，元の応答の MBR スコアが著しく低いということはモデルが生成した候補群の多くが正解から逸脱した内容に収束しており，類似度に基づいた MBR による選好が機能しない可能性が高いと考えられるからである．

手法 2 (MBR-DPO) プロンプト x_i に対する応答候補集合の中で MBR スコアが最も高い応答を正例 y_i^+，MBR スコアが最も低い応答を負例 y_i^- としてデータセット $\{x_i, y_i^+, y_i^-\}$ を構築し，DPO を行う．ただし，手法 1 と同様に，元の応答の MBR スコアの順位 r が事前に定めた閾値を下回る場合には，元の応答を正例とし，MBR スコアが最も低い応答を負例とする．このとき，元の応答の MBR スコアが最も低かったことで正例と負例が一致してしまう場合，2 番目に MBR スコアが低い応答を負例として用いる．

4 実験

Step1: ベースラインモデルの構築 llm-jp-3-980m[1] と中間学習済みの SmolLM3-3B-Base [14] に対し，既存のデータセットを用いて事後学習を施したモデルをそれぞれベースラインモデルとした．llm-jp-3-980m の学習には llm-jp/instruct3[2] を，SmolLM3-3B-Base の学習には huggingface/alignment-handbook/recipes/smollm3[3] を参考にした．

Step2: 応答生成 ベースラインモデルの構築時に使用した各データについて，最後の assistant 応答直前までの系列をプロンプトとした．このプロンプトをベースラインモデルに入力して，新たな assistant 応答を複数生成した．本実験では，計算コストと候補多様性のバランスを考慮し，各プロンプトにつき 50 個の応答を生成した．

Step3: 応答選好 応答候補集合の各要素をベースラインモデルに入力して，各応答候補の埋め込みベクトルを取得し，式 2 に基づいて MBR スコアを算出した．算出された MBR スコアに基づき，前節の手法に従って応答を選好し，新たな事後学習用データセットを作成した．手法 1 および手法 2 における「元の応答を採用する閾値」として元の応答の順位が下位 0%，30%，50%，70%，100%に含まれている場合の 5 通りを設定し，それぞれの場合についてデータセットを作成した．なお，下位 0%の場合は常にモデルの選好に従う設定であり，下位 100%の場合は常に元の応答を採用する設定となる．

Step4: 再学習 作成したデータセットを用いて再度事後学習を行った．再学習の際には，応答が単語のみのように短いデータセットおよび Chain-of-Thought (CoT) を要する (think タグを含む) データセット，Tool の使用を要する (tool タグを含む) データセットを除外した．また，元の応答や生成された応答が空文字列であったデータに関しても同様に除外した．

Step5: 評価 llm-jp-3 の性能評価には llm-jp-eval [15][4] を，SmolLM3 の性能評価には Lighteval [16] を用いた．llm-jp-eval においては，機械翻訳，コード生成，要約タスクを除いた．SmolLM3 の評価については，think 機能を無効にし，Lighteval の評価指標として，指示追従能力の評価には IFEval，算術的な推論能力の評価には GSM8K，多分野にわたる知識・推論能力の評価には MMLU を採用した．

5 実験結果および考察

本節では，単一モデルによる自己生成・自己選好データを用いた事後学習の有効性について，llm-jp-3 および SmolLM3 の各モデルに対する実験結果を基に考察する．llm-jp-3 の結果を表 1 に，SmolLM3 の結果を表 2 に示す．表中の値は複数タスクの結果を平均したものであり，タスクごとの評価結果は付録 A に示した．

llm-jp-3 の結果 手法 1 では，閾値 0%，30%，50%の設定でベースラインを上回り，特に閾値 30%で最も高い性能を示した．一方で，閾値 100%の設定では性能が低下した．この結果は，元のデータで再度学習するよりも，自己生成・自己選好データを適切な割合で導入することで性能向上に寄与することを示している．手法 2 では，一部タスクで性能向上は見られたものの平均スコアは一貫してベースラインを下回る結果となった．特に，閾値 0%の設定における平均スコアの大幅な低下は，モデルの選好のみに依存することのリスクを示唆している．しかし，閾値を設けて元の応答を活用することで性能低

1) https://huggingface.co/llm-jp/llm-jp-3-980m
2) https://github.com/llm-jp/instruct3/tree/main
3) https://github.com/huggingface/alignment-handbook/tree/main/recipes/smollm3
4) 本研究では v2.1.1 を使用した．

下が抑制されたことから，閾値設定が一種の安全装置として機能したといえる．総じて，llm-jp-3 では，DPO よりも SFT の枠組みにおいて，自己生成・自己選好データの有効性が確認された．

SmolLM3 の結果　手法 1 では，閾値 0%および 100%の極端な設定でベースラインを下回ったのに対し，閾値 30%および 70%の中間的な設定ではスコアが向上した．これは llm-jp-3 と同様に，自己生成データと元の高品質データを適切に混合することの有効性を示唆している．手法 2 では，算術的推論を要するタスクでの性能低下は見られたものの，多くの設定で性能向上が確認された．特に指示追従タスクでの性能向上が顕著であった．SmolLM3 においては llm-jp-3 とは対照的に，自己生成・自己選好データを用いた DPO が SFT よりも比較的有効に機能した．

総合的分析　2 つの結果を通して，モデルやタスクによって異なる傾向が見られた．llm-jp-3 では手法 1 が，SmolLM3 では手法 2 が比較的有効であるという結果であった．この差異は，ベースラインモデルの初期能力や学習データの性質，タスク特性の違いに起因すると考えられる．しかし，共通する傾向として，閾値 0%や 100%といった極端な設定よりも，中間的な閾値設定が高い性能を示した．これは，MBR スコアに基づく選好によって自己生成データを導入しつつ，閾値設定に基づいて元の応答のスコアが低い場合には元の応答を保持するという本手法の枠組みが有効であったことを示している．この枠組みは，モデルが生成する誤ったコンセンサスを学習するリスクを抑制し，自己生成データによる学習効果を最大化する上で重要な役割を担うことを示唆する．

なお，本研究で応答選好の基準として用いた MBR スコアは応答候補間の埋め込みベクトルによる意味的類似度に基づいた指標である．そのため，ベースラインモデルの埋め込み性能が低い場合や応答候補に意味的な多様性がない場合にはうまく機能しない可能性がある．また，この指標が必ずしも全てのタスクに対して有効であるとは限らない．そのため，生成応答の多様性確保やタスクに対する選好指標の有効性の検証は今後の課題である．

6　おわりに

本研究では，既存データセットの応答部分を学習対象の言語モデル自身で生成・選好することにより，新たな事後学習用データセットを構築し，再度学習を行う自己完結型の事後学習手法について検討した．特に，外部モデルや人手による評価に依存せず，単一の言語モデルのみを用いて自己生成および自己選好を行う点に着目し，Minimum Bayes Risk (MBR) に基づく選好基準を導入した．

実験の結果，汎化性能の向上は限定的であったが，一部の設定では性能向上が見られた．これにより，単一モデルによる自己生成・自己選好データを用いた事後学習が，適切な学習枠組みと組み合わせることで有効に機能する可能性が示唆された．

今後の展望としては，MBR スコアの算出におけるより効果的な自己評価指標の検討や，タスクに対する選好指標の有効性の検証，応答候補の多様性を明示的に制御した生成手法の導入などが挙げられる．

表 1　llm-jp-eval での評価時の平均スコア．括弧内の値は当該指標におけるベースラインとの差分（Δ）である．

手法	閾値 [%]	平均 [%]
ベースライン		19.51
手法 1 (MBR-SFT)	0	19.83 (+**0.32**)
	30	20.15 (+**0.64**)
	50	19.52 (+**0.01**)
	70	19.25 (−0.26)
	100	19.02 (−0.49)
手法 2 (MBR-DPO)	0	16.05 (−3.46)
	30	18.85 (−0.66)
	50	18.85 (−0.66)
	70	19.21 (−0.30)
	100	18.78 (−0.73)

表 2　Lighteval での評価時の平均スコア．括弧内の値は当該指標におけるベースラインとの差分（Δ）である．

手法	閾値 [%]	平均 [%]
ベースライン		65.79
手法 1 (MBR-SFT)	0	65.29 (−0.50)
	30	65.93 (+**0.14**)
	50	65.57 (−0.22)
	70	66.05 (+**0.26**)
	100	65.36 (−0.43)
手法 2 (MBR-DPO)	0	66.06 (+**0.27**)
	30	67.17 (+**1.38**)
	50	65.59 (−0.20)
	70	67.72 (+**1.93**)
	100	67.27 (+**1.48**)

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research)，および，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の助成を受けたものです．本研究成果の一部は，九州大学情報基盤研究開発センター研究用計算機システムの「一般利用」，および，「ABCI 3.0 開発加速利用」の支援を受けて産総研及び AIST Solutions が提供する ABCI 3.0 を利用して得られたものです．

参考文献

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

[2] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[3] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[4] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[5] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[7] Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 391–398, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Ian Wu, Patrick Fernandes, Amanda Bertsch, Seungone Kim, Sina Khoshfetrat Pakazad, and Graham Neubig. Better instruction-following through minimum bayes risk. In **The Thirteenth International Conference on Learning Representations**, 2025.

[9] Shankar Kumar and William Byrne. Minimum Bayes-risk word alignments of bilingual texts. In **Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)**, pp. 140–147. Association for Computational Linguistics, July 2002.

[10] Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In **Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004**, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[11] António Farinhas, José de Souza, and Andre Martins. An empirical study of translation hypothesis ensembling with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11956–11970, Singapore, December 2023. Association for Computational Linguistics.

[12] Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation, 2023.

[13] Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 4265–4293, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

[15] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会発表論文集, pp. 2085–2089, 2024.

[16] Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023.

A　タスクごとの評価結果

llm-jp-eval のタスクごとの評価結果を表 3 に，Lighteval のタスクごとの評価結果を表 4 に示す．

表 3　llm-jp-eval の評価結果．括弧内の値は当該指標におけるベースラインとの差分（Δ）である．

手法	閾値 [%]	平均	NLI	QA	RC	CR	HE-JA	HE-EN	EL	FA	MR	IF	BBH
ベースライン		19.51	34.10	13.02	34.38	35.23	20.76	20.55	11.08	4.88	1.90	17.94	20.82
手法 1	0	19.83 (+**0.32**)	35.30 (+**1.20**)	11.80 (−1.22)	36.72 (+**2.34**)	35.52 (+**0.29**)	17.78 (−2.98)	20.72 (+**0.17**)	17.08 (+**6.00**)	5.07 (+**0.19**)	1.94 (+**0.04**)	16.60 (−1.34)	19.64 (−1.18)
	30	20.15 (+**0.64**)	34.60 (+**0.50**)	13.70 (+**0.68**)	37.08 (+**2.70**)	35.04 (−0.19)	18.94 (−1.82)	18.96 (−1.59)	17.06 (+**5.98**)	5.20 (+**0.32**)	2.04 (+**0.14**)	17.79 (−0.15)	21.23 (+**0.41**)
	50	19.52 (+**0.01**)	34.15 (+**0.05**)	13.98 (+**0.96**)	37.19 (+**2.81**)	35.53 (+**0.30**)	17.34 (−3.42)	16.56 (−3.99)	15.18 (+**4.10**)	5.01 (+**0.13**)	1.90 (±0.00)	16.83 (−1.11)	21.06 (+**0.24**)
	70	19.25 (−0.26)	33.76 (−0.34)	13.65 (+**0.63**)	37.19 (+**2.81**)	33.97 (−1.26)	17.12 (−3.64)	17.51 (−3.04)	15.30 (+**4.22**)	5.27 (+**0.39**)	1.01 (−0.89)	16.19 (−1.75)	20.78 (−0.04)
	100	19.02 (−0.49)	34.50 (+**0.40**)	13.08 (+**0.06**)	35.97 (+**1.59**)	35.19 (−0.04)	17.04 (−3.72)	16.92 (−3.63)	14.02 (+**2.94**)	4.80 (−0.08)	1.36 (−0.54)	15.35 (−2.59)	21.00 (+**0.18**)
手法 2	0	16.05 (−3.46)	24.23 (−9.87)	9.80 (−3.22)	20.46 (−13.92)	34.65 (−0.58)	20.71 (−0.05)	20.83 (+**0.28**)	3.17 (−7.91)	5.20 (+**0.32**)	1.19 (−0.71)	15.62 (−2.32)	20.65 (−0.17)
	30	18.85 (−0.66)	29.04 (−5.06)	13.98 (+**0.96**)	31.90 (−2.48)	35.06 (−0.17)	21.37 (+**0.61**)	19.87 (−0.68)	11.79 (+**0.71**)	4.77 (−0.11)	1.90 (±0.00)	16.01 (−1.93)	21.65 (+**0.83**)
	50	18.85 (−0.66)	30.10 (−4.00)	13.30 (+**0.28**)	33.50 (−0.88)	35.04 (−0.19)	20.18 (−0.58)	21.14 (+**0.59**)	11.38 (+**0.30**)	4.45 (−0.43)	2.39 (+**0.49**)	14.62 (−3.32)	21.24 (+**0.42**)
	70	19.21 (−0.30)	30.69 (−3.41)	12.67 (−0.35)	34.02 (−0.36)	34.79 (−0.44)	20.75 (−0.01)	20.00 (−0.55)	13.13 (+**2.05**)	4.68 (−0.20)	2.11 (+**0.21**)	17.96 (+**0.02**)	20.51 (−0.31)
	100	18.78 (−0.73)	32.95 (−1.15)	14.01 (+**0.99**)	36.11 (+**1.73**)	33.97 (−1.26)	14.57 (−6.19)	18.83 (−1.72)	13.78 (+**2.70**)	4.26 (−0.62)	1.86 (−0.04)	15.47 (−2.47)	20.82 (±0.00)

表 4　Lighteval の評価結果．括弧内の値は当該指標におけるベースラインとの差分（Δ）である．

			GSM8K	MMLU	IFEval			
手法	閾値 [%]	平均	extractive match	exact match	prompt level strict acc	inst level strict acc	prompt level loose acc	inst level loose acc
ベースライン		65.79	54.81	57.68	64.51	73.26	67.84	76.62
手法 1	0	65.29 (−0.50)	54.28 (−0.53)	56.90 (−0.78)	63.77 (−0.74)	73.14 (−0.12)	67.28 (−0.56)	76.38 (−0.24)
	30	65.93 (+**0.14**)	56.25 (+**1.44**)	58.15 (+**0.47**)	64.14 (−0.37)	73.38 (+**0.12**)	67.28 (−0.56)	76.38 (−0.24)
	50	65.57 (−0.22)	53.90 (−0.91)	57.77 (+**0.09**)	64.33 (−0.18)	73.50 (+**0.24**)	67.28 (−0.56)	76.62 (±0.00)
	70	66.05 (+**0.26**)	55.50 (+**0.69**)	58.40 (+**0.72**)	64.88 (+**0.37**)	73.38 (+**0.12**)	68.02 (+**0.18**)	76.14 (−0.48)
	100	65.36 (−0.43)	53.37 (−1.44)	57.59 (−0.09)	64.14 (−0.37)	73.50 (+**0.24**)	67.28 (−0.56)	76.26 (−0.36)
手法 2	0	66.06 (+**0.27**)	52.69 (−2.12)	60.21 (+**2.53**)	64.51 (±0.00)	74.10 (+**0.84**)	67.84 (±0.00)	76.98 (+**0.36**)
	30	67.17 (+**1.38**)	53.22 (−1.59)	59.62 (+**1.94**)	67.10 (+**2.59**)	75.66 (+**2.40**)	69.50 (+**1.66**)	77.94 (+**1.32**)
	50	65.59 (−0.20)	52.24 (−2.57)	54.54 (−3.14)	65.62 (+**1.11**)	74.82 (+**1.56**)	68.76 (+**0.92**)	77.58 (+**0.96**)
	70	67.72 (+**1.93**)	54.06 (−0.75)	59.22 (+**1.54**)	67.47 (+**2.96**)	76.38 (+**3.12**)	70.43 (+**2.59**)	78.78 (+**2.16**)
	100	67.27 (+**1.48**)	53.83 (−0.98)	58.54 (+**0.86**)	66.91 (+**2.40**)	75.78 (+**2.52**)	70.24 (+**2.40**)	78.30 (+**1.68**)