対照的デコーディングを用いた指示学習データの合成

一瀬 達矢 [1]　Youmi Ma[1]　大井 聖也 [1]　小池 隆斗 [1]　岡崎 直観 [1,2,3]
[1] 東京科学大学　[2] 産業技術総合研究所　[3]NII LLMC
{tatsuya.ichinose,masanari.ohi,ryuto.koike}@nlp.comp.isct.ac.jp
{ma.y,okazaki}@comp.isct.ac.jp

概要

大規模言語モデル（LLM）によって指示学習データの応答文を合成する試みは，人手に頼らず低コストに大量のデータを合成できる点で有望である．既存手法の多くは高性能な指示学習済みモデルの応答文をそのまま学習データに利用している．ところが，生成される応答文には事前学習などによって獲得された特徴が反映されている可能性があり，指示学習の目的に最適化されているとは限らない．本研究では，指示学習前後のモデルを用いた対照的デコーディングによって，指示学習に有用な応答文を生成する手法 CoDIT を提案する．複数のベンチマーク・モデルを用いた実験の結果，CoDIT は既存研究よりも高い性能改善を達成した．

1　はじめに

ユーザの指示に適切に応答させるために，事前学習済みモデルを指示文と応答文のペアで追加学習する指示学習が広く用いられている [1]．指示学習データを人手で構築することはコストと時間がかかるため，大規模化しにくいという欠点がある [2]．そこで，高性能な指示学習済み LLM を教師モデルとして用い，あらかじめ用意した指示文に対する応答文を生成することで，大量の高品質データを高速に合成するアプローチが登場している [3, 4, 5, 6]．

指示学習データを合成する既存手法は，教師モデルの応答を学習データとしてそのまま用いるが，これは指示学習の目的と整合しない可能性がある．先行研究において，指示学習は事前学習で獲得した知識を指示に応じて適切に引き出すことを学ぶ段階であると言われている [7, 8, 9]．しかし，事前学習で獲得された知識はモデルごとに異なる．教師モデルの応答は事前学習で獲得した知識を色濃く反映するため，応答をそのまま指示学習データとして採用すると，教師とは異なる生徒モデルの指示学習に向かない可能性がある．

本研究では，学習効果の高い指示学習データを合成する手法 **CoDIT** (**Co**ntrastive **D**ecoding for **I**nstruction-**T**uning Dataset) を提案する．**CoDIT** の概要を図 1 に示す．CoDIT では，指示学習前後のモデルが出力するトークンの対数確率の差を用いる対照的デコーディング [10] で文生成（合成）を行う．これにより，モデルが事前学習で獲得した能力と，指示学習で追加的に獲得した能力を分離し，後者を強調する形で応答文を合成することを狙う．結果として，指示学習で獲得すべき能力を反映したデータを効果的に抽出でき，指示学習効果が高いデータセットを構築できると期待できる．

実験では，複数のオープン・ウェイトモデルを教師として提案手法でデータセットを構築し，その学習効果を分析する．その結果，モデルの規模やアーキテクチャに依らず，単に指示学習済みモデルの応答文を用いる手法よりも，提案手法は異なる生徒モデルの指示学習後の性能を一貫して向上させることが分かった．さらに，先行研究で合成された指示学習データセットと比較した場合においても，本研究のデータセットを用いた方が複数のベンチマークで一貫して高い性能を示した．その高い指示学習効果が確認されたため，今後のモデル開発に資する有用な資源として利用できるよう，構築したデータセットを公開した[1)]．

2　指示学習データ合成手法 CoDIT

本研究では，所与の指示文に対する応答文を合成するモデルを「教師モデル」，その合成データを用いて指示学習を行うモデルを「生徒モデル」と呼ぶ．提案手法 CoDIT では，教師モデルによる応答生成において，対照的デコーディング [10] を指示学習前後のモデルに適用する．

1) https://huggingface.co/datasets/Tatsuya-Ichinose/CoDIT

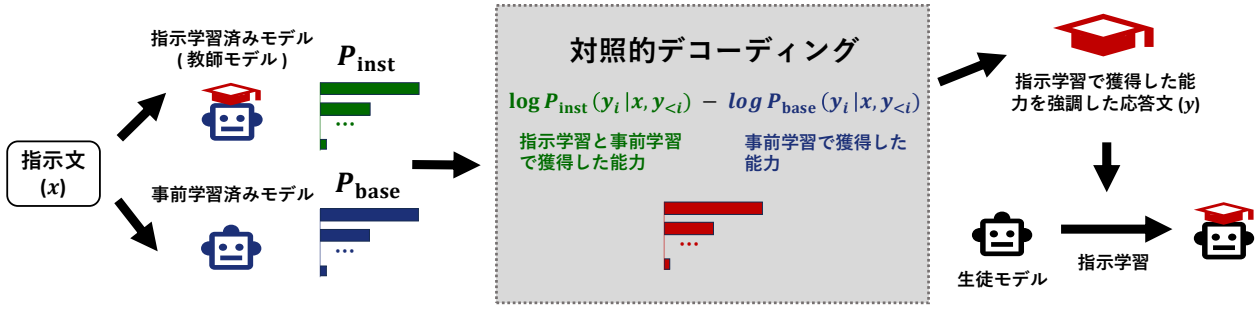

図 1　指示学習前後のモデルを用いた対照的デコーディングにより，指示学習で獲得した能力を強調した応答文を合成．

具体的には，入力される指示文を x，生成される応答文を y とする．教師モデルが応答文の i 番目のトークン y_i を生成する際，以下の式 (1) に示すように，指示学習済みモデル (inst) と事前学習済みモデル (base) の対数確率の差 $s(y_i; x, y_{<i})$ を計算し，この値が高いトークンをサンプリングする．

$$s(y_i; x, y_{<i}) = \log P_{\text{inst}}(y_i \mid x, y_{<i}) - \log P_{\text{base}}(y_i \mid x, y_{<i}) \tag{1}$$

生成されたトークン y_i は，これまでの生成系列 $y_{<i}$ の末尾に追加され，次のトークン y_{i+1} の生成に用いられる．この過程を繰り返すことで，最終的な応答文を合成する．

しかし，単純に対数確率の差のみを用いると，事前学習済みモデルの確率が極端に低く不適切なトークンが過大評価されたり，逆に双方が高い確率を計算するトークンが過小評価される問題が生じる．その結果，対照的デコーディングを語彙全体に適用すると，文脈上妥当なトークンが抑制され，意味の通らない文が生成されてしまう恐れがある．そこで本研究では，先行研究 [10] を踏まえ，式 (2) で定義される妥当性制約を導入する．

$$\mathcal{V}_{\text{head}}(x, y_{<i}) = \left\{ v \in \mathcal{V} : P_{\text{inst}}(v \mid x, y_{<i}) \geq \alpha \max_{w \in \mathcal{V}} P_{\text{inst}}(w \mid x, y_{<i}) \right\} \tag{2}$$

ここで $\mathcal{V}$ は全語彙集合，$\alpha \in [0, 1]$ は制約の厳しさを制御するハイパーパラメータである．本制約は，指示学習済みモデルが比較的高い確率を割り当てるトークン集合 $\mathcal{V}_{\text{head}}$ にのみ生成候補を絞ることで，生成文の自然さを損なうことなく対照的デコーディングの適用を可能にする．

以上を踏まえると，本研究では，妥当性制約を考慮したスコア $s'(y_i; x, y_{<i})$ を式 (3) のように定義し，対照的デコーディングを行う．

$$s'(y_i; x, y_{<i}) = \begin{cases} s(y_i; x, y_{<i}) & y_i \in \mathcal{V}_{\text{head}}(x, y_{<i}) \\ -\infty & \text{それ以外} \end{cases} \tag{3}$$

式 (3) により，指示学習後のモデルが生成しやすく，指示学習前のモデルが生成しにくいトークンが多く含まれるデータセットを構築できる．

3　実験

3.1　実験設定

教師モデル　本研究では，教師モデルとして Qwen3-8B[2]，Qwen3-30B-A3B[3]，および gemma-3-27b-it[4] を採用した．選定理由は主に以下の 3 点である．(1) 事前・指示学習済みモデルの双方が公開され，CoDIT の対照的デコーディングが適用可能であること，(2) 異なるモデル系列や規模を含めることで多角的な検証が可能であること，(3) 高品質な合成データを生成するために十分な汎用対話能力を有していること．なお，生成時の設定は付録 A に記載する．

生徒モデル　本研究では，生徒モデルとして Llama-3.1-8B[5]，Qwen3-8B-Base[6]，および gemma-3-4b-pt[7] を採用した．選定理由は，異なるモデル系列・規模において提案手法の有効性を広範に検証するためである．なお，学習設定は付録 B に記載する．

2) https://huggingface.co/Qwen/Qwen3-8B
3) https://huggingface.co/Qwen/Qwen3-30B-A3B
4) https://huggingface.co/google/gemma-3-27b-it
5) https://huggingface.co/meta-llama/Llama-3.1-8B
6) https://huggingface.co/Qwen/Qwen3-8B-Base
7) https://huggingface.co/google/gemma-3-4b-pt

表 1　提案手法およびベースラインで作成した指示学習用データセットで学習したモデルに対する，各指示追従能力の評価スコア．

教師モデル	手法	生徒モデル: Llama-3.1-8B				生徒モデル: Qwen3-8B-Base				生徒モデル: gemma-3-4b-pt			
		Wild Bench	Alpaca Eval2.0		MT Bench	Wild Bench	Alpaca Eval2.0		MT Bench	Wild Bench	Alpaca Eval2.0		MT Bench
		WB-Score	LC(%)	WR(%)	AVG	WB-Score	LC(%)	WR(%)	AVG	WB-Score	LC(%)	WR(%)	AVG
Qwen3-8B	Baseline	43.5	**47.92**	52.55	7.31	60.1	53.46	64.76	8.44	30.2	29.25	40.42	6.39
	CoDIT	**48.5**	44.07	**58.73**	**7.54**	**63.0**	**56.83**	**69.45**	**8.53**	**34.7**	**33.43**	**47.18**	**6.50**
Qwen3 -30B-A3B	Baseline	43.0	42.03	51.78	7.28	57.6	53.10	59.47	**8.44**	31.4	32.47	40.36	5.87
	CoDIT	**48.2**	**46.94**	**56.76**	**7.37**	**60.4**	**55.22**	**63.79**	8.42	**34.1**	**35.50**	**43.46**	**6.28**
gemma-3 -27b-it	Baseline	47.4	41.43	69.51	**7.44**	59.6	46.23	71.67	**8.36**	34.1	33.04	55.12	**6.79**
	CoDIT	**55.0**	**52.17**	**75.95**	7.43	**62.9**	**51.10**	**76.26**	8.31	**38.9**	**34.57**	**59.8**	6.71

データセットの合成　本研究では LMSYS-Chat-1M [5] に含まれる指示文に対して CoDIT を適用して応答文を合成する．LMSYS-Chat-1M は Chatbot Arena 等のウェブサイトから人間と LLM の対話履歴を 100 万件収集したデータセットである．本研究では，先行研究 [6] に倣い，重複除去，テンプレート形式の指示文の除去，および個人情報を含む指示文の除去を行い，250,333 件の英語の指示文を利用した．これらの指示文を入力として CoDIT を適用し，対応する応答文を生成することで，データセットを構築した．

α のチューニング　本研究では，MT-Bench [11] を用いて対照的デコーディングのハイパーパラメータ α をチューニングした．α に関する詳細な実験設定および結果は，付録 C に示す．

評価データセットと評価指標　合成したデータセットによる指示学習の効果を検証するため，WildBench [12]，AlpacaEval 2.0 [13]，MT-Bench [11] の 3 つのベンチマークを用いた．

MT-Bench は，80 件の高品質な 2 ターン対話タスクから構成され，GPT-4o（2024-08-06）を用いた LLM-as-a-Judge により，各タスクを 10 点満点で評価する．同じ指示に対して 5 回生成させて評価を行い，その平均を最終スコアとする．

AlpacaEval 2.0 は，805 件の指示に対する LLM の応答を，参照モデルである GPT-4 Turbo（1106）の応答とペアワイズ比較するベンチマークである．評価には，GPT-4 Turbo（1106）を用いて算出される勝率（Win Rate; WR）を用いる．加えて，応答長バイアスを補正した Length-Controlled Win Rate（LC） [14] も用いる．

WildBench は，MT-Bench における評価問題の少なさや，AlpacaEval におけるタスク難易度の偏りといった既存ベンチマークの課題に対処することを目的として設計された，より包括的な評価ベンチマークである．実世界の対話ログから収集された 1024 件のタスクから構成され，最大 5 ターンの対話履歴を含む．本研究では，GPT-4o（2024-05-13）による 10 点満点評価に基づき，スコア 5 を基準として再スケーリングした WB-Score を評価指標として用いる．

3.2　実験結果

表 1 に各設定における評価結果を示す．上段はベースラインである指示学習後のモデルのみを用いて生成した応答文で学習した結果（Baseline）を，下段は CoDIT によって得られた応答文で学習した結果を示している．本実験で教師モデル・生徒モデルとして用いた各モデルの評価結果は，付録 D に記載する．

提案手法の一貫した性能向上　多くの設定において，提案手法はベースラインと比較して一貫して高い性能を示していることが分かる．具体的には，WildBench における WB-Score は平均で 4.31 ポイント向上し，AlpacaEval 2.0 においても LC および WR がそれぞれ平均で 3.43 ポイント，5.08 ポイント改善している．一方，MT-Bench ではモデル間の性能差は限定的であり，提案手法によりスコアが低下する組み合わせも観測された．詳細な解析結果は付録 E に示す．

モデルごとの分析　生徒モデルごとの WindBench のスコアに着目すると，ベースラインと比較して，CoDIT は Llama-3.1-8B （平均 ≈ +5.93）および gemma-3-4b-pt （≈ +4.00）で大きな改善が確認されたのに対し，Qwen3-8B-Base （≈ +3.01）では改善幅

表 2　指示学習用データセットに関する先行研究との性能比較．提案手法により構築したデータセットは，すべてのベンチマークにおいて先行研究を上回る性能を示す．

データセット	件数	指示文	応答文	生徒モデル: Llama-3.1-8B				生徒モデル: Qwen3-8B-Base			
				Wild Bench	Alpaca Eval2.0		MT Bench	Wild Bench	Alpaca Eval2.0		MT Bench
				WB-Score	LC(%)	WR(%)	AVG	WB-Score	LC(%)	WR(%)	AVG
WildChat [4]	442,172	Human	GPT-3.5, GPT-4	20.7	13.25	8.59	6.43	31.4	18.66	12.78	7.17
Llama-3.1-LMSYS-Chat-1M-Synth [6]	453,737	Human	Llama-3.1-405B-Instruct	30.6	23.07	23.76	7.21	42.4	28.54	26.26	8.01
Gemma-2-LMSYS-Chat-1M-Synth [6]	453,861	Human	gemma-2-27b-it	35.7	35.47	27.44	6.95	46.3	42.91	31.1	7.46
CoDIT-Gemma3	250,333	Human	gemma-3-27b-it	**55.0**	**52.17**	**75.95**	7.43	62.9	51.10	**76.26**	8.31
CoDIT-Qwen3-8B	250,333	Human	Qwen3-8B	48.5	44.07	58.73	**7.54**	**63.0**	**56.83**	69.45	**8.53**
CoDIT-Qwen3-30B	250,333	Human	Qwen3-30B-A3B	48.2	46.94	56.76	7.37	60.4	55.22	63.79	8.42

が相対的に小さい．この傾向は，Qwen3-8B-Base がベースライン時点で既に高いスコアを示しており，改善余地が限定的であったと考えられる．また，教師モデル別の WindBench のスコアにおいては，gemma-3-27b-it（≈ +5.23），Qwen3-8B（≈ +4.13），Qwen3-30B-A3B（≈ +3.58）と，いずれの教師モデルにおいても改善が一貫して観測された．

α のチューニング　α に関する WildBench のアブレーション実験の結果，いずれの α の設定においても，ベースラインと比較して一貫して高い性能が得られることが確認された．図 2 は，gemma-3-27b-it を教師モデル，Llama-3.1-8B を生徒モデルとし，$\alpha = 0.01, 0.04, 0.07, 0.1$ の各設定で CoDIT により構築したデータセットについて，ベースライン（指示学習済みモデルのみを用いて構築したデータセット）との WildBench スコアを比較した結果を示している．これらの結果から，本研究で示した性能向上は，MT-Bench において最も高いスコアを与える α を選択したことによるものではなく，α の値に大きく依存せず一貫した性能改善をもたらす提案手法自体の有効性に起因するものであると結論づけられる．

先行研究との比較　CoDIT を用いて生成した指示学習データセットをモデル開発に利用することの有用性を検証するため，先行研究である Llama-3.1-LMSYS-Chat-1M-Synth，Gemma-2-LMSYS-Chat-1M-Synth，および WildChat を用いた場合と学習効果を比較した．表 2 にその結果を示す．いずれのベンチマークにおいても，提案手法により構築したデータセットを用いたモデルは，先行研究に基づくデータセットを用いた場合と比較して，一貫して高い性能を示すことが確認された．この結果は，CoDIT で合成されたデータセットが，既存の指示学習データセットと比較してより有用な学習資源であることを示している．なお，構築したデータセットは公開しており[1)]，学習設定の詳細は付録 B に記載する．

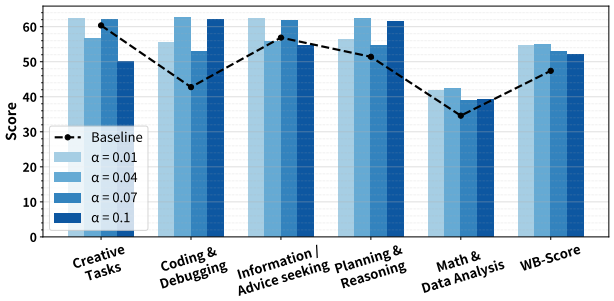

図 2　異なる α における WildBench のアブレーション結果（教師モデル：gemma-3-27b-it，生徒モデル：Llama-3.1-8B）

4　おわりに

本研究では，指示学習前後のモデルを用いた対照的デコーディングにより，指示学習の目的に最適化した応答文を合成し，学習効果を高める手法 CoDIT を提案した．複数の教師モデルを用いてデータセットを構築し，対照的デコーディングの有無による比較を行った結果，異なる生徒モデルの指示学習後の性能を，複数のベンチマークで一貫して向上させることが分かった．さらに，先行研究に基づく指示学習データセットとの比較を通じて，本手法により構築したデータセットが有用な学習資源であることを示した．今後の課題としては，指示学習後に適用される強化学習の影響や，推論型モデルに対する本手法の有効性について検証したい．

謝辞

本研究は JSPS 科研費 25H01137 の助成を受けたものです．本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究（22501）により得られたものです．本研究は，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施しました．

参考文献

[1] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **The Tenth International Conference on Learning Representations**, 2022.

[2] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data. In **First Conference on Language Modeling**, 2024.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023. https://lmsys.org/blog/2023-03-30-vicuna/.

[4] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In **The Twelfth International Conference on Learning Representations**, 2024.

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In **The Twelfth International Conference on Learning Representations**, 2024.

[6] Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. Building instruction-tuning datasets from human-written instructions with open-weight large language models. In **Second Conference on Language Modeling**, 2025.

[7] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[8] Arnav Gudibande, Eric Wallace, Charlie Victor Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[9] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In **The Twelfth International Conference on Learning Representations**, 2024.

[10] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive Decoding: Open-ended text generation as optimization. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2023.

[11] 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.

[12] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In **The Thirteenth International Conference on Learning Representations**, 2025.

[13] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[14] Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-Controlled AlpacaEval: A simple debiasing of automatic evaluators. In **First Conference on Language Modeling**, 2024.

[15] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In **The International Conference for High Performance Computing, Networking, Storage and Analysis**, 2020.

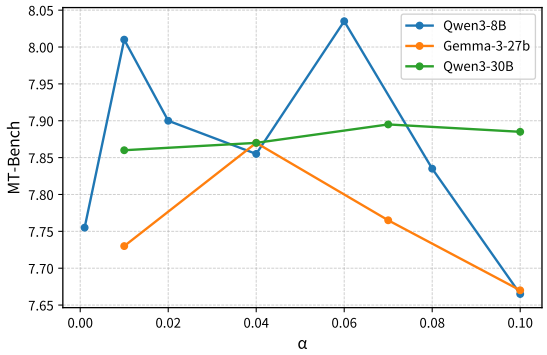

図 3　α に対する MT-Bench スコア（2 モデル平均）の変化

表 3　本研究における使用モデル群のベンチマーク評価

モデル	Wild Bench	Alpaca Eval2.0		MT Bench
	WB-Score	LC(%)	WR(%)	AVG
Llama-3.1-8B-Instruct[8]	34.5	22.47	23.30	7.41
gemma3-4b-it[9]	47.2	43.11	70.51	7.84
Qwen3-8B	59.6	57.72	62.14	8.28
Qwen3-30B-A3B	63.2	58.92	66.61	8.62
gemma-3-27b-it	64.1	65.61	84.11	8.81

A　生成に用いた設定

生成時のハイパーパラメータは全実験で共通とし，Temperature は 1.0，Top-p は 1.0，最大生成トークン数は 4096 とした．

B　学習に用いた設定

学習はすべて 4 枚の NVIDIA H100 SXM5 を用いて実施し，DeepSpeed ZeRO [15] によるメモリ最適化を適用した．学習に用いたハイパーパラメータは，特に断りのない限り，先行研究 [6] に従って設定した．また，学習スクリプトは LLM-jp が公開している指示学習用リポジトリを参考に実装した[10]．

学習率は最大値を 2.5×10^{-5}，最小学習率を 2.5×10^{-6} とし，Warmup 比率 0.1 の Cosine スケジューラを適用した．最適化手法には AdamW を用い，$\beta_1 = 0.90$，$\beta_2 = 0.95$ とした．学習エポック数は 2，有効バッチサイズは 512 に設定した．ただし，gemma-3-4b-pt では上記設定のままでは性能低下が確認されたため，MT-Bench を指標として学習率探索を行い，最大学習率を 1.0×10^{-5}，最小学習率を 1.0×10^{-6} に変更した．

さらに，Qwen3-30B-A3B を教師モデルとして合成したデータセットを用いて gemma-3-4b-pt を学習した際，メモリ不足が発生したため，応答長が 100,000 文字を超える 3 件のデータについては，応答長が 100,000 文字となるよう切り詰めた．また，WildChat データセットについては英語の指示文のみを使用し，複数ターン対話は最初のターンのみを学習に用いた．加えて，学習時にメモリ不足が生じたため，指示文が 4K 文字以上，または応答文が 8K 文字以上のデータを除外し，合計 36,326 件のデータを削除した．

8）https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

9）https://huggingface.co/google/gemma-3-4b-it

10）https://github.com/llm-jp/llm-jp-sft

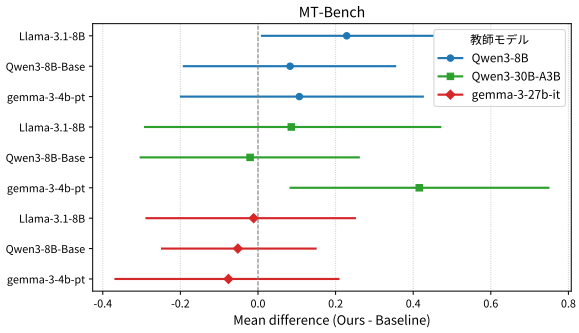

図 4　MT-Bench における平均差と 95% 信頼区間

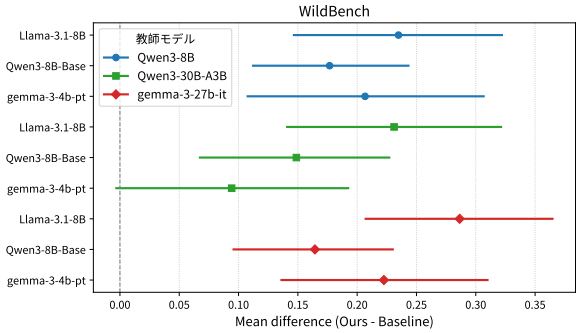

図 5　WildBench における平均差と 95% 信頼区間

C　α のチューニング

本研究では，式 (2) のハイパーパラメータ α を決定するため，MT-Bench を検証データとして用いた．複数の α に対して対照的デコーディングにより学習データを構築し，Qwen3-8B-Base および Llama-3.1-8B を指示学習した後，両モデルにおける MT-Bench スコアの平均が最大となる α を選択した．図 3 に，α と MT-Bench スコア（2 モデル平均）の関係を示す．その結果，教師モデルが Qwen3-8B の場合は $\alpha = 0.06$，gemma-3-27b-it では $\alpha = 0.04$，Qwen3-30B-A3B の場合は $\alpha = 0.07$ が最適であり，本研究ではこれらの値を採用した．

D　使用したモデル群の評価結果

表 3 に参考値として，本研究の生徒モデルに対応する公式の指示学習済みモデル，および教師モデルの評価結果を掲載した．

E　MT-Bench における検出力分析

図 4 は MT-Bench，図 5 は WildBench における教師・生徒モデル組み合わせごとの平均差（Ours − Baseline）と 95% 信頼区間を示す．MT-Bench（80 問）では，各タスクカテゴリ内で質問を，各質問内で 5 回実行された評価結果を復元抽出する階層ブートストラップを 10,000 回行い，スコア差の 95% 信頼区間を推定した．WildBench（1024 問）では単純ブートストラップを 10,000 回行い，同様に信頼区間を算出した．なお，WildBench のスコアにはスケーリング前の 10 点満点評価値を用いている．

MT-Bench では，平均差が負となる組み合わせも観測されるが，95% 信頼区間全体が負となる例はなく，提案手法による性能低下は統計的に有意とは言えない．