大規模言語モデルによる自己説明の忠実性は改善するか？

土井 智暉 [1]　磯沼 大 [1,2,3]　谷中 瞳 [1]
[1] 東京大学　[2] エディンバラ大学　[3] 国立情報学研究所
{doi-tomoki701, hyanaka}@is.s.u-tokyo.ac.jp　m.isonuma@ed.ac.uk

概要

大規模言語モデルによる自己説明は，ブラックボックスなモデルの挙動を解釈可能な表現に変換することが期待できる．しかし近年，自己説明が必ずしもモデルの挙動を忠実に反映しておらず、自己説明と実際の挙動が矛盾しうることが明らかになっている．本研究では，入力文中で予測に最も影響を与える語を特定し，これを教師信号として継続学習することによって，自己説明の忠実性が改善するかを検証する．実験により，複数の分類タスクおよび説明様式の条件下で継続学習の有効性を検証し，とくに未学習の分類タスクおよび説明様式においても忠実性が改善することを確認した．

1　はじめに

大規模言語モデル（LLM）は高度な言語生成能力を有し，自身の挙動の過程や根拠の説明（自己説明）を生成する能力をも備えている．このような自己説明は，本来ブラックボックスである LLM の挙動を人間に解釈可能な表現に変換する試みとして注目されている [1]．

しかし，近年の研究により，LLM の自己説明が実際の挙動を忠実に反映していないことが指摘されている．具体的には，自己説明においてモデルが有する社会的バイアスが隠蔽される [2, 3]，自己説明に従い入力を編集しても期待された挙動が得られない [1, 4]，挙動に影響を及ぼしたと推定された重要内容が自己説明に反映されない [1, 4, 5]，などの課題が確認されている．自己説明は高い説明性を提供し得る一方で，前述のような非忠実な自己説明はモデルに関する誤った理解を助長する危険性があり，忠実な自己説明能力の確立が求められている．

本研究では，自己説明の忠実性の改善に向けた継続学習のフレームワークとして「自己説明学習」を提案する．自己説明学習では，忠実な自己説明を生成するタスク（自己説明タスク）を，入力文章中で予測に最も影響を与える語を特定するタスクとして定式化する．その上で，モデルが既に学習したデータに自己説明タスクのデータを混ぜて継続学習を行うことで，自己説明の忠実性改善を図る．

図 1: 自己説明タスクにおける学習データと評価データのイメージ．本研究では，未学習の分類タスクおよび説明様式での自己説明タスク（青背景）への汎化の有無についても検証する．

実験では，指示チューニング済み LLM（7B，13B）について，複数の分類タスク（感情分析，自然言語推論，短文読解の多選択肢問題）および説明様式（1語，2 語）の条件下で，自己説明学習の有効性を検証する．実験の結果，学習データと同一設定での自己説明だけでなく，未学習の分類タスクや説明様式での自己説明においても，忠実性が改善することを確認した（図 1）．

2　問題設定

本研究では，自己説明タスクの一つとして，モデルの予測に最も影響を与えた入力文章中の語を特定するタスクを考える．具体的には，あるタスクについて入力文章 $x = (w_1, w_2, \ldots w_n)$ が与えられたときに，モデル θ の予測が $\hat{y}$ であったとする（$\hat{y} = \arg\max_y p_\theta(y \mid x)$）．このとき，入力文章中の語で，入力文章から除いたときに $\hat{y}$ の予測確率を最も減少させる語を，予測に最も影響を与えた語 w^* として定義する．

$$w^* = \arg\max_{w \in x} p_\theta(\hat{y} \mid x) - p_\theta(\hat{y} \mid x_{-w}) \quad (1)$$

ただし，x_{-w} は文章 x から語 w を除いた文章である．語をそのまま除くと文章の文法構造が崩れるため，先行研究 [4] を踏襲し，本研究では “[REDACTED]” という文字列で置き換える．本研究では，自己説明タスクの学習により，モデルの予測に最も影響を与えた語を特定する正確性が向上するかを検証する．

3 提案手法

本研究では，モデルの忠実性を向上させるためのフレームワーク「自己説明学習」を提案する．自己説明学習は，モデル特有の自己説明データセットの構築と，これを活用した継続学習という二段階から構成される．

3.1 データセットの構築

任意の分類タスク T における入力文章の集合を $X_T = \{x_1, x_2, \ldots x_m\}$，自己説明学習の対象モデルを θ とおく．全ての入力文章 x について，モデル θ の予測に最も影響を与える語 w^*（式 1）を特定し，自己説明データセット $SE_T = \{(x_1, w_1^*), (x_2, w_2^*), \ldots (x_m, w_m^*)\}$ を構築する．

3.2 継続学習

モデル θ を用いて構築した自己説明データセット SE_T を活用し，同モデル θ の継続学習を行う．継続学習では，学習前に有していた能力が学習後に損なわれる破滅的忘却 [6] を軽減するために，モデル θ の指示チューニングデータも併用する [7]．

4 実験

本章では，指示チューニング済みモデルに自己説明学習を行い，その有効性を検証する．評価にあたっては，未学習の分類タスクや説明様式（予測に最も影響を与えた「2 語」）における性能の評価も行い，自己説明学習の汎化可能性を議論する．

4.1 モデル

本実験では，LLM として Tulu-2[1] の 7B および 13B モデルを用いる．Tulu-2 はオープンソースの指示チューニング済みモデルであり [8]，Llama-2 [9] をベースモデルとしている．自己説明学習の効果を検証するために，後述の分類タスクについて 50,000 件からなる自己説明データセットを構築し，Low-Rank Adaptation (LoRA，[10]) を用いて 1 エポックの継続学習を行う[2]．このとき，指示チューニングデータとして，Tulu-2 の指示チューニング時に使用されたデータ[3]のうち，ランダムにサンプリングされた 10,000 件を併用する．ベースラインとしては，継続学習なしのオリジナルの Tulu-2，指示チューニングデータのみで継続学習した Tulu-2，そして，自己説明タスクで要求されている語数（1 つあるいは 2 つ）の語を文章中からランダムにサンプリングする Random モデルを用いる．

4.2 データセット

自己説明タスクを学習・評価する分類タスクのデータセットとして，感情分析タスクから Sentiment140 [11]，自然言語推論タスクから SNLI [12]，多選択形式の短文読解問題から babiQA [13] を用いる．なお，分類タスクおよび自己説明タスクで使用するプロンプトについては付録 A.1 に示す．

学習データ　学習用データセットの構築にあたっては，Sentiment140 と SNLI からそれぞれ 50,000 件ずつ入力文をランダムにサンプリングする．これらを用いて，学習を行う各モデル（Tulu-2 7B, 13B）について，予測に最も影響を与える 1 語を事例ごとに特定し，学習用の自己説明データセット $SE_{\text{Sent}}, SE_{\text{SNLI}}$ を構築する．

評価データ　評価用データセットの構築にあたっては，Sentiment140，SNLI，babiQA から学習データと重複しないようにそれぞれ 1,000 件ずつ入力文をランダムにサンプリングする．そして，評価を行う継続学習前後の各モデルについて，予測に最も影響を与える「1 語」，および「2 語」を事例ごとに特定し，評価用の自己説明データセットをそれぞれ構築する．ここで「予測に最も影響を与える 2 語」は，入力文章 x 中の 2 語で，同時に “[REDACTED]” で置き換えたときに元の予測 $\hat{y}$ の予測確率を最も減少させるような 2 語 (w_1^*, w_2^*) として定義する．なお，Random モデルの評価用データセットについては，継続学習前の Tulu-2 7B のものを使用する．

4.3 評価

本研究では，自己説明タスクでの正解率に基づいてモデルの自己説明忠実性を評価する（2 章参照）．以下の 2 種類の指標を導入する．

1） https://huggingface.co/collections/allenai/tulu-v2-suite-6551b56e743e6349aab45101

2） ハイパーパラメータの詳細は付録 A.2 に示す．

3） https://huggingface.co/datasets/allenai/tulu-v2-sft-mixture

表 1: 継続学習前後のモデルの自己説明タスクにおける性能．No-Train は継続学習しなかった場合の性能，D_{Inst} は指示チューニングデータのみで継続学習した場合の性能を表す．$D_{\text{Inst}} + SE_{Sent}$，$D_{\text{Inst}} + SE_{SNLI}$ はそれぞれ Sentiment140, SNLI で自己説明学習を行った場合の性能を表す．

継続学習設定	Sentiment140				SNLI				babiQA			
	1 語		2 語		1 語		2 語		1 語		2 語	
	Weight	Flip	Weight	Flip	Weight	Flip	Weight	Flip	Weight	Flip	Weight	Flip
Tulu-2 7B												
No-Train	0.245	0.189	0.068	0.066	0.227	0.228	0.065	0.055	**0.713**	**0.750**	0.061	0.053
D_{Inst}	0.236	0.204	0.080	0.074	0.185	0.182	0.065	0.059	0.664	0.690	**0.072**	**0.066**
$\boldsymbol{D_{\text{Inst}}+SE_{\text{Sent}}}$	**0.581**	**0.491**	**0.146**	**0.128**	0.347	0.314	0.049	0.046	0.254	0.198	0.059	0.058
$\boldsymbol{D_{\text{Inst}}+SE_{\text{SNLI}}}$	0.367	0.301	0.075	0.072	**0.639**	**0.559**	**0.134**	**0.124**	0.116	0.057	0.033	0.034
Tulu-2 13B												
No-Train	0.433	0.344	0.064	0.060	0.370	0.322	0.051	0.046	**0.632**	0.616	0.081	0.076
D_{Inst}	0.414	0.331	0.074	0.071	0.341	0.289	0.059	0.058	0.583	0.565	**0.082**	**0.082**
$\boldsymbol{D_{\text{Inst}}+SE_{\text{Sent}}}$	**0.615**	**0.497**	**0.130**	**0.111**	0.419	0.375	0.090	0.089	0.528	0.507	0.067	0.065
$\boldsymbol{D_{\text{Inst}}+SE_{\text{SNLI}}}$	0.415	0.343	0.063	0.073	**0.722**	**0.677**	**0.197**	**0.192**	0.625	**0.621**	0.080	0.080
Random	0.125	0.111	0.068	0.064	0.083	0.072	0.027	0.029	0.093	0.093	0.027	0.025

Weighted Accuracy　ある語が予測に与える影響が大きければ大きいほど，モデルの自己説明はその語を忠実に挙げることがより強く求められる．そこで，自己説明タスクの正解データ w^*（式 1）について，予測に対する影響の大きさ $I_\theta(w^* \mid x)$ で重みづけした正解率 Weighted Accuracy (Weight) を導入する $(I_\theta(w^* \mid x) = p_\theta(\hat{y} \mid x) - p_\theta(\hat{y} \mid x_{-w^*}))$.

$$\text{Weight} = \frac{\sum_x I_\theta(w^* \mid x) \cdot \mathbb{1}(\hat{w} = w^*)}{\sum_x I_\theta(w^* \mid x)} \tag{2}$$

$\hat{w}$ はモデルが自己説明で挙げた語（1 語あるいは 2 語）を表す．なお，正解データとモデルの回答間の一致については，語数および全ての語についての語幹が一致している場合に限り一致と判定する．

Flip-case Accuracy　一方で，正解データ w^* が予測に与える影響が小さい場合であっても，入力文章における w^* の有無がモデルの予測ラベル $\hat{y}$ を変化させる事例においては，依然として自己説明が w^* を反映することが強く要求される．そのような予測ラベルが変化する事例 x_{flip} のみに基づいて計算された正解率 Flip-case Accuracy (Flip) を導入する．

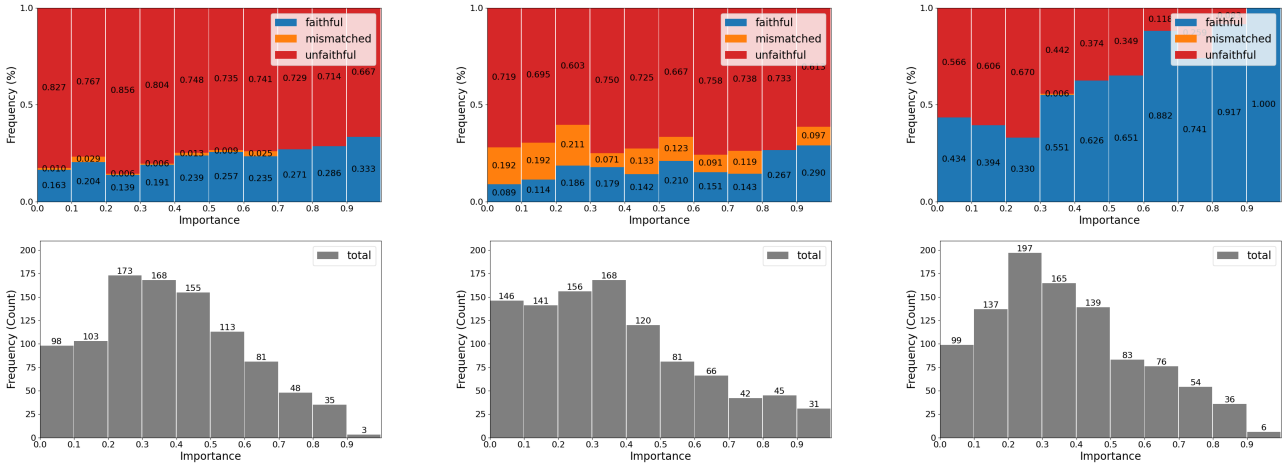
$$\text{Flip} = \frac{\sum_{x_{flip}} \mathbb{1}(\hat{w} = w^*)}{\sum_{x_{flip}} 1} \tag{3}$$

5　結果

学習データと同一設定での性能　自己説明タスクにおける評価結果を表 1 に示す．自己説明学習後のモデルは，学習データと同一の分類タスクおよび説明様式（1 語）において，一貫して学習前よりも高い性能を示している．例えば，SNLI における 1 語での自己説明タスクにおいて，元の Tulu-2 7B（No-Train）は Weight で 0.227，Flip で 0.228 であるのに対し，自己説明学習後のモデル（$D_{\text{Inst}} + SE_{\text{SNLI}}$）はそれぞれ 0.639，0.559 となり，いずれも大きく改善している．改善傾向の詳細を分析するために，継続学習前後の 7B モデルについて，正解データ w^* が予測に与える影響の大きさ別に自己説明タスクの性能を示したものが 図 2 である．この図から，自己説明学習後のモデル（c）では，とりわけ忠実な自己説明が要求される事例（とくに $I_\theta(w^* \mid x)$ が大きい事例のこと，4.3 節参照）で顕著な性能の向上が確認できる．

未学習の分類タスクへの汎化　表 1 から，自己説明学習後のモデルは，学習データに含まれていない分類タスクの一部においても自己説明タスクの性能改善を示していることがわかる．例えば，感情分析（Sentiment140）で自己説明学習を行った 7B モデル（$D_{\text{Inst}} + SE_{\text{Sent}}$）は，自然言語推論（SNLI）の自己説明タスクにおいても Weight で 0.12，Flip で 0.08，それぞれ性能が改善している．感情分析と自然言語推論については必要とされる推論が互いに異なる分類タスクであり，自己説明タスクにおける正解データの分布も異なると考えられる．それにもかかわらず性能改善が確認されたことは、自己説明学習がタスク間の汎化可能性を有していることを示唆している．

未学習の説明様式への汎化　表 1 から，自己説明学習後のモデルは，学習データに含まれていない 2



(a) No-Train　(b) D_{Inst}　(c) D_{Inst} + SE_{SNLI}

図 2: SNLI での継続学習前後の Tulu-2 7B における，正解データの予測に与える影響の大きさ Importance ($I_\theta(w^* \mid x)$) 別の自己説明タスクの性能（上段）と事例数（下段）．“faithful”，“mismatched”，“unfaithful” はモデルが自己説明で挙げた語に割り当てられたカテゴリであり，それぞれ正解データの語と一致，文章中のいずれの語とも不一致，正解データ以外の語と一致を意味する（$I_\theta(w^* \mid x)$ および一致判定の詳細は 4.3 節参照）．

語による自己説明タスクにおいても，学習前と比較して性能を改善させていることがわかる．この結果は，自己説明学習が説明様式間で汎化する可能性を示唆している．このことから，例えば句や文，さらには自由記述といったより複雑で表現力の高い説明様式においても，1 語あるいは 2 語といった比較的単純な説明様式で自己説明学習を行うことで，その忠実性を向上させることが期待できる．

分類の正誤と自己説明の忠実性との関係　ここでは，分類タスクでの正誤が自己説明タスクでの性能と関係しているかを分析する[4]．表 2 は，SNLI において，分類タスクでの正誤によって事例を分割し，正誤事例別に自己説明タスクでの性能を評価した結果を示している．いずれのモデルも，分類タスクで誤答した事例に比べて正答した事例の方が，自己説明タスクでの性能が高いことがわかる．自己説明タスクでの性能は，説明対象とする分類タスクでの性能と関係していると考えられる．ただし，自己説明学習前のモデルは分類タスクで正答した事例においても忠実性が低いことを述べておく．SNLI での 7B モデルにおいては，ランダムモデルと比較して自己説明タスクの性能差が 0.1~0.2 程度に留まっている．

表 2: SNLI における分類タスクでの正誤事例別の自己説明タスクの性能．

継続学習設定	分類正答事例		分類誤答事例	
	1 語		1 語	
	Weight	Flip	Weight	Flip
Tulu-2 7B				
No-Train	0.237	0.256	0.188	0.179
D_{Inst}	0.201	0.212	0.152	0.157
$\boldsymbol{D_{\text{Inst}}+SE_{\text{Sent}}}$	0.385	0.398	0.242	0.219
$\boldsymbol{D_{\text{Inst}}+SE_{\text{SNLI}}}$	**0.722**	**0.790**	**0.448**	**0.342**
Tulu-2 13B				
No-Train	0.386	0.359	0.266	0.222
D_{Inst}	0.379	0.371	0.210	0.162
$\boldsymbol{D_{\text{Inst}}+SE_{\text{Sent}}}$	0.448	0.416	0.291	0.273
$\boldsymbol{D_{\text{Inst}}+SE_{\text{SNLI}}}$	**0.810**	**0.806**	**0.333**	**0.271**
Random	0.089	0.085	0.061	0.047

6　おわりに

本研究では，LLM の自身の挙動に対する自己説明の忠実性が，予測に最も影響を与える語を教師信号とした継続学習によって改善できるのかを検証した．実験では指示チューニング済みモデル Tulu-2 に継続学習を適用し，複数の分類タスクおよび説明様式の条件下で有効性を評価した．結果として，学習時の自己説明タスクにおける忠実性の改善に加え，一部の未学習の分類タスクおよび説明様式においても忠実性が改善することを確認した．本研究が自己説明による LLM の説明性向上に寄与し，より信頼性の高い LLM の開発に資することを期待する．

4)　分類タスクでの性能については付録 A.3 に示す．

謝辞

本研究の一部は JST さきがけ JPMJPR21C8，JSPS 科研費学術変革領域研究（B）「ナラティブ意識学」JP24H00809 の支援を受けたものである．

参考文献

[1] Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, 2023.

[2] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In **Advances in Neural Information Processing Systems**, 2023.

[3] Katie Matton and Robert Ness & Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In **ICLR 2024 Workshop on Secure and Trustworthy Large Language Models**, 2024.

[4] Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In **Findings of the Association for Computational Linguistics: ACL 2024**, 2024.

[5] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2024.

[6] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. **arXiv preprint arXiv:2308.08747**, 2023.

[7] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, 2022.

[8] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2. **arXiv preprint arXiv:2311.10702**, 2023.

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[11] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. **CS224N project report, Stanford**, Vol. 1, No. 12, p. 2009, 2009.

[12] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, 2015.

[13] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. **arXiv preprint arXiv:1502.05698**, 2015.

表 3: 実験に使用したプロンプトのテンプレート．{text},{premise},{hypothesis} などの波括弧で示された部分は，事例ごとに適切な文字列で置き換えられる．

データセット	タスク	プロンプト
Sentiment140	感情分析	Text: {text}\n\n Question: What would you classify the sentiment of the text as? The text can contain redacted words marked with [REDACTED]. Answer either 'Positive' or 'Negative' in a single word. Do not explain the answer.
	自己説明（1 語）	Text: {text}\n\n Question: List the single most important word for determining the sentiment of the text, such that without this word the sentiment cannot be determined. Answer one word following 'Answer:'. Do not explain the answer.
	自己説明（2 語）	Text: {text}\n\n Question: Identify a pair of words from the text that together are the most important for determining its sentiment, such that without these words, the sentiment cannot be determined. Answer in the JSON format: {"Pair": ["word1", "word2"]}
SNLI	自然言語推論	Sentence: {premise}\n\n Question: Does this sentence imply that '{hypothesis}'? The sentence can contain redacted words marked with [REDACTED]. Answer either 'Yes', 'No', or 'Maybe' in a single word. Do not explain the answer.
	自己説明（1 語）	Sentence: {premise}\n\n Question: List the single most important word in the sentence, for determining if this sentence imply that '{hypothesis}'. Answer one word following 'Answer:'. Do not explain the answer.
	自己説明（2 語）	Sentence: {premise}\n\n Question: Identify a pair of words from the sentence that together are the most important for determining if this sentence imply that '{hypothesis}'. Answer in the JSON format: {"Pair": ["word1", "word2"]}
babiQA	短文読解 QA	Context: {context}\n\n Question: {question} The answer choices are {options} The context can contain redacted words marked with [REDACTED]. Answer either '(A)', '(B)', or '(C)' in a single word. Do not explain the answer.
	自己説明（1 語）	Context: {context}\n\n Question: List the single most important word in the context, for answering the question: '{question}' Answer one word following 'Answer:'. Do not explain the answer.
	自己説明（2 語）	Context: {context}\n\n Question: Identify a pair of words from the context that together are the most important, for answering the question: '{question}' Answer in the JSON format: {"Pair": ["word1", "word2"]}

A 付録

A.1 プロンプトの詳細

実験で使用したプロンプトのテンプレートを表 3 に示す．プロンプトの設計は先行研究 [4] に基づく．

A.2 継続学習時のハイパーパラメータ

Tulu-2 7B および 13B の継続学習時のハイパーパラメータについては，学習率を 1e-4，LoRA ランクを 64，α を 16，ドロップアウトを 0.1，エポック数を 1 とした．

A.3 分類タスクでの性能

継続学習前後の Tulu-2 7B および 13B の各分類タスクでの性能を表 4 に示す．

表 4: 継続学習前後のモデルの各分類タスクでの正解率．Sentimet140 は 2 値分類タスク，SNLI および babiQA は 3 値分類タスクとなっている．

継続学習設定	Sentiment140	SNLI	babiQA
Tulu-2 7B			
No-Train	0.750	**0.778**	0.952
D_{Inst}	0.812	0.671	0.946
$\boldsymbol{D_{\text{Inst}}}$**+**$\boldsymbol{SE_{\text{Sent}}}$	**0.828**	0.674	**0.955**
$\boldsymbol{D_{\text{Inst}}}$**+**$\boldsymbol{SE_{\text{SNLI}}}$	0.809	0.657	0.947
Tulu-2 13B			
No-Train	0.718	**0.831**	0.968
D_{Inst}	0.814	0.722	**0.973**
$\boldsymbol{D_{\text{Inst}}}$**+**$\boldsymbol{SE_{\text{Sent}}}$	**0.816**	0.700	0.964
$\boldsymbol{D_{\text{Inst}}}$**+**$\boldsymbol{SE_{\text{SNLI}}}$	0.786	0.646	0.967