LLM のクロスリンガル知識編集に関する分析

友成光 森下皓文 角掛正弥 今一修 十河泰弘
株式会社日立製作所 研究開発グループ
hikaru.tomonari.oj@hitachi.com

概要

本研究では，クロスリンガル知識編集における言語間転移と誤った知識の上書きがどの程度生じるかを明らかにするため，新たにデータセットを構築し，評価実験を行った．具体的には，Wikidata から多言語の知識トリプルを収集し，反実仮想の目的語を導入することで，モデルがもともと持つ知識に左右されずに編集効果を測定できる仕組みを整えた．さらに，新指標「過剰適応耐性（OF-TOL）」を提案し，編集対象外の主語に対する誤った知識の上書きを定量化した．Llama-3-8B モデルを用いた検証では，述語の種類や更新パラメータによって言語間転移の度合いが大きく変動する一方，既存知識が上書きされる副作用が一貫して生じることが分かった．

1 はじめに

大規模言語モデル（LLM）が多種多様な自然言語処理タスクで高い性能を示す一方，世界の動的な変化に伴い，モデル内部の古い知識を修正したり，新しい知識を取り入れることが求められている [1, 2, 3]．こうした問題に対し，既存のモデルを局所的に更新できる手法として知識編集 [4, 5, 6, 7] が注目されている．知識編集の主な目的は，モデルが学習済みの知識のうち特定の要素のみを改変または忘却し，その他の知識への影響を最小限に抑えることである．このように必要な修正を局所的かつ迅速に施しながら，モデル全体の性能を維持できる点が評価され，LLM を長期的かつ柔軟に活用するためのアプローチとして期待が高まっている．

一方，LLaMA [8] や GPT-4 [9] などのマルチリンガル LLM は，多言語環境での利用を想定して設計されているが，ある言語で編集を施したモデルが他の言語においてどのように振る舞うかは，いまだ十分に解明されていない．近年では，知識編集をモノリンガルで検討するだけでなく，編集を実行するソース言語と編集結果を評価するターゲット言語が異なる場合を対象とした「クロスリンガル知識編集」の研究も報告されている [10, 11, 12]．たとえば，英語の質問応答データセットである ZsRE[13] を中国語に翻訳した評価データを用いてマルチリンガル LLM を編集した結果，ソース言語側で編集した知識がターゲット言語にも転移する可能性が示唆されている [10]．しかし，各種評価指標の精度は十分に高いとは言えず，言語間での知識転移のメカニズム解明やモデル性能の改善に向けた研究は，いまだ道半ばである．

本研究ではクロスリンガル知識編集の評価を行うため，(1) 主語・述語・目的語に多言語対訳が存在し，(2)LLM が既知の知識として保持していないトリプルを含むデータセットを構築する．具体的には，Wikidata[14] から多言語の知識トリプルを抽出し，その目的語を別の候補に置き換えることで反実仮想データセットを生成する．この手法により，モデルが本来持つ知識では答えられない状況を作り出し，クロスリンガル知識編集の効果を適切に評価可能となる．さらに，既存の局所性（Locality）指標 [15, 16, 17] を拡張し，編集対象の知識と強く関連する要素のみを定量的に評価する新たな指標を提案する．従来の局所性指標は，編集対象と無関係な知識を編集前後で保持できるかを主に測定していたが，関連性の高い知識を上書きする可能性までは考慮していなかった．そこで本研究では，関連性の高い要素に着目した分析を通じ，LLM における知識の獲得および忘却のメカニズムを多言語環境下でより正確に理解することを目指す．

述語ごとに知識トリプルを分類し，パラメータを変化させた実験では，条件によって最大 22%の知識転移が観測されたものの，多くの設定で転移率は低水準にとどまった．また，新たな知識を追加するすべてのケースで，関連性の高い既存知識が上書きされる現象が確認された．これにより，クロスリンガル編集は一定の効果を示しつつも，既存知識の維持が難しいことが示唆された．

2 手法

本研究では，まず単言語環境における知識編集手法を基盤とし，それを多言語環境へ拡張することで，ソース言語で編集した知識がターゲット言語へどの程度転移するかを検証する枠組みを構築する．以下では，単言語環境での知識編集の基本的な問題設定から説明し，次いでクロスリンガル知識編集の定義を示す．その後，知識編集の評価指標と評価実験に用いるデータセットの作成手順を述べる．

2.1 知識編集の問題設定

編集前のモデルを p_θ，モデルに与える入力を x，モデルの出力を y，編集対象を y_e としたとき，知識編集は編集後のモデル $p_{\theta'}$ が，以下のような性質を満たすことを目的とする:

$$\arg\max_y p_{\theta'}(y \mid x) = \begin{cases} y_e, & x \in \mathfrak{X}_e, \\ \arg\max_y p_\theta(y \mid x), & x \notin \mathfrak{X}_e. \end{cases} \tag{1}$$

編集後のモデル $p_{\theta'}$ は，ある知識を問う同義のテキスト集合 $\mathfrak{X}_e$ について，$x \in \mathfrak{X}_e$ に対しては y_e を応答し，それ以外の入力に対しては元のモデル p_θ と同様の出力を生成することを期待する．

本研究では，反実仮想の目的語を含むトリプル (s, r, o') を編集情報として用いる．たとえば，

$$(s, r, o') = (\text{日本}, \text{首都}, \text{パリ})$$

のように目的語を事実と異なる候補に差し替えることで，モデルが保持している既存知識に左右されにくい形で編集の効果を測定可能にする．編集の際は，(s, r) を結合してモデルに入力し，目的語 o' を出力するよう，モデルパラメータを θ から θ' に更新する．

2.2 クロスリンガル知識編集の問題設定

クロスリンガル知識編集はソース言語における知識編集の結果，ソース言語と異なるターゲット言語について知識が転移しているかを評価する．編集前のマルチリンガルモデルを $p_{m\theta}$，ターゲット言語の入力を x^t とし，ソース言語からターゲット言語への翻訳関数を $I^t(\cdot)$ とすると，クロスリンガル知識編集は編集後のマルチリンガル言語モデル $p_{m\theta'}$ が，以下のような性質を満たすことを目的とする:

$$\arg\max_y p_{m\theta'}(y \mid x^t) = \begin{cases} I^t(y_e^s), & x^t \in I^t(\mathfrak{X}_e^s), \\ \arg\max_y p_{m\theta}(y \mid x^t), & x^t \notin I^t(\mathfrak{X}_e^s). \end{cases} \tag{2}$$

式 (2) は，ソース言語側で編集された知識が，翻訳関数 $I^t(\cdot)$ を介してターゲット言語の同義表現にも適用されることを示している．

ここで，ソース言語の編集対象トリプルを $(s_{\text{src}}, r_{\text{src}}, o'_{\text{src}})$ とし，そのターゲット言語版を $(s_{\text{tgt}}, r_{\text{tgt}}, o'_{\text{tgt}}) = I^t(s_{\text{src}}, r_{\text{src}}, o'_{\text{src}})$ と定義する．まずソース言語側で $(s_{\text{src}}, r_{\text{src}}, o'_{\text{src}})$ を用いてモデルを編集し，編集後のモデル $p_{m\theta'}$ にターゲット言語で $(s_{\text{tgt}}, r_{\text{tgt}})$ を入力した際に o'_{tgt} が正しく生成されるかを確認することで，言語間の知識転移を評価する．

2.3 評価指標

ターゲット正解率 (Target Accuracy, T-ACC) 学習した知識トリプル $(s_{\text{src}}, r_{\text{src}}, o'_{\text{src}})$ をもとに，ターゲット言語で $(s_{\text{tgt}}, r_{\text{tgt}})$ をモデルに入力し，期待する目的語 o'_{tgt} を **exact match** で生成できる割合を

$$\text{T-ACC} = \frac{\#(\text{正解 } o'_{\text{tgt}} \text{ を生成})}{\#(\text{総サンプル})} \tag{3}$$

で定義する．T-ACC が高いほど，ソース言語の知識編集がターゲット言語でも転移し，正答を生成できていることを示す．なお，知識編集前は反実仮想知識の生成は不可能であるため，T-ACC の値は 0.0 である．

過剰適応耐性 (Over-Fitting Tolerance , OF-TOL) ソース言語で $(s_{\text{src}}, r_{\text{src}}, o'_{\text{src}})$ を学習後，ターゲット言語において主語をランダムに変えた $(s_{\text{tgt}}^{\text{rand}}, r_{\text{tgt}})$ に対し，誤って o'_{tgt} を生成してしまう割合を「過剰適応率 (OF)」とする．これを

$$\text{OF} = \frac{\#(\text{ランダム主語で } o'_{\text{tgt}} \text{ を生成})}{\#(\text{総サンプル})}, \quad \text{OF-TOL} = 1 - \text{OF} \tag{4}$$

と定義する．OF-TOL が高いほど，編集対象外の主語への影響が小さく，特定の主語・述語のみが正しく更新されていることを意味する．なお，知識編集前は編集対象外の主語に与える影響はないため，OF-TOL の値は 1.0 である．

2.4 損失値による補足的評価

T-ACC と OF-TOL は，モデルが実際に生成するテキストを観測することで評価を行う．一方，モデルの予測トークン確率の変化などの詳細な変化をテキスト生成結果だけでは検知することができない．

そこで本研究では，同一モデル内での評価を補完する目的として，モデルが正解応答（ターゲット言語に翻訳された目的語）をどの程度高確率で生成す

るかを，交差エントロピー損失（以下，損失値）で測定する．これにより，生成結果には表れなかったモデルの予測トークン確率を考慮することが可能となる．

損失値 L は，モデルの生成確率 $p_{\theta'}$ に基づいて $L = -\log p_{\theta'}\left(I^t(y_e^s) \,\middle|\, x^t\right)$ で計算される．損失値 L が低いほど，モデルが正解応答を高確率で生成できていることを示す．

2.5 多言語反実仮想データセットの構築

本研究では，LLM のクロスリンガル知識編集能力を評価するにあたり，既存の知識で既に正解可能なトリプルでは知識編集の効果を正確に測定することが困難であるという課題に対処するため，反実仮想データセットを構築する．

反実仮想の知識トリプルは Wikidata[14] から SPARQL クエリを用いて抽出する[1)]．Wikidata は「主語-述語-目的語」の形式で世界中の知識を統合的に管理しており，多言語対応も充実している．データ作成手順は以下の通りである．

1. 述語の一覧を英語（EN），日本語（JA），中国語（ZH）の組み合わせで取得し，ID や URL，その他の識別子を除外する．
2. 各述語について，主語と目的語の組み合わせを最大 10 万件まで取得する．
3. 取得した知識トリプルの中から，異なる言語間で文字列が完全一致するような目的語（例：数値など）を除外する．
4. 主語および目的語が他のトリプルと重複しないようにフィルタリングし，知識トリプルを選定する．これは出現頻度の高い目的語への偏りを防ぐためである．得られた知識トリプルの例は付録 A に示す．
5. 目的語をランダムに入れ替えた知識トリプル (s, r, o') を生成する．反実仮想の目的語を設定することにより，LLM が有する既存知識に依存せずクロスリンガル知識編集の評価を行うことが可能となる．
6. 主語を，元の主語 s や入れ替えた目的語 o' の正しい主語ではない他の主語 s' にランダムに入れ替えた知識トリプル (s^{rand}, r, o') を生成する．これは前述の式 (4) による過剰適応の評価に使用する．

1） https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/ja

本研究ではこうして得られた知識トリプルの中から，述語「首都，建築様式，動脈支配，作曲者，監督」に関して「動脈支配」は 83 サンプル，その他はそれぞれ 100 サンプルを用いた．[2)]

3 実験

本研究の目的は，クロスリンガル知識編集における知識転移と，過剰適応について調べることである．具体的には，述語の種類と更新するパラメータについて，知識の転移および忘却に与える影響を分析した．

3.1 実験設定

本実験では Llama-3-8B[3)] を用いて，異なる述語・更新パラメータについてクロスリンガル知識編集を実施した．知識編集時のパラメータは付録 B に示す．実験は各データサンプルを用いて 5 回（5 エポック）モデルパラメータの更新を行い，生成結果の評価を行うことをすべてのサンプルについて繰り返した．

また，ターゲット言語で評価を行うときに知識トリプルと無関係な後続文や知識編集したソース言語の目的語が出力される場合が多くあったため，ターゲット言語の one-shot 例を与えることでターゲット言語で目的語が生成されるように促した．

3.2 実験結果

本研究では，Llama-3-8B モデルを用い，異なる述語および更新パラメータを設定してクロスリンガル知識編集を行った．その結果を表 1 および図 1，図 2 に示す．ここでは，転移率（T-ACC）の向上と，既存知識の上書きをどの程度抑えられるか（OF-TOL）の観点から結果を議論する．

述語の違いによる影響　表 1 は，ソース言語とターゲット言語の組み合わせ，および述語の種類を変えた場合の T-ACC と OF-TOL を示す．まずソース言語とターゲット言語が同一のとき，T-ACC がほぼ 1.0 に到達する一方で OF-TOL は 0.00 ~ 0.46 程度まで低下しており，新規に獲得した知識が関連性のある他の主語にも波及してしまうことが確認できる．一方，異なる言語間で編集を行った場合は，T-ACC が 0.00 ~ 0.22 にとどまり，クロスリンガルの知識転

2） 述語「首都」の知識トリプルは次の wikipedia 記事から取得した．https://en.wikipedia.org/wiki/List_of_national_capitals

3） https://huggingface.co/meta-llama/Meta-Llama-3-8B

表 1　異なる述語に関するクロスリンガル知識編集（5 エポック）の結果．#1 は T-ACC，#2 は OF-TOL を表す．各述語についての知識トリプルとして，動脈支配は 83 サンプル，それ以外は 100 サンプルを用いた．

		EN		JA		ZH	
述語	src	#1	#2	#1	#2	#1	#2
首都	EN	1.00	0.13	0.22	0.97	0.15	0.99
	JA	0.04	0.99	0.99	0.07	0.08	0.95
	ZH	0.02	1.00	0.09	0.96	1.00	0.10
建築様式	EN	0.95	0.11	0.07	0.96	0.03	0.97
	JA	0.07	0.99	0.99	0.01	0.11	0.89
	ZH	0.07	0.98	0.12	0.87	1.00	0.01
動脈支配	EN	0.78	0.22	0.02	0.99	0.00	1.00
	JA	0.05	0.99	1.00	0.00	0.17	0.84
	ZH	0.05	0.94	0.19	0.80	1.00	0.01
作曲者	EN	1.00	0.23	0.05	0.99	0.00	1.00
	JA	0.00	1.00	0.98	0.01	0.06	0.94
	ZH	0.02	1.00	0.08	0.93	0.97	0.04
監督	EN	0.98	0.46	0.01	0.99	0.00	1.00
	JA	0.02	1.00	1.00	0.00	0.08	0.95
	ZH	0.00	1.00	0.09	0.93	1.00	0.01

移は限定的な結果となった．ただし「首都」のような有名な固有名詞を扱う述語では，他の述語よりも転移がやや高いと同時に，OF-TOL の悪化が比較的小さい傾向にある．これは，知識トリプルの対訳関係が比較的明確な場合に知識転移しやすいことを示唆する．

パラメータ更新領域の違いによる影響　図 1 は述語「首都」を対象に，更新対象のパラメータを変化させた際の T-ACC と OF-TOL をエポックごとに示す．付録 C に示すモデルの更新パラメータが多いほど T-ACC は顕著に向上するが，他方で OF-TOL が大きく低下するケースが多かった．これは，大きなモデル変更ほど編集対象以外の知識にも影響が及ぶリスクが高いことを示している．一方，Attention 層のみの更新のように，変更範囲を狭めれば過剰な忘却は起こりにくい一方，新規知識の獲得自体が困難になり，転移率も低めにとどまる傾向が見られた．

損失値評価からの補足的知見　図 2 では，英語をソースとし，日本語をターゲットに選んで編集した場合の目的語の損失値を比較した．編集対象の主語と目的語の組み合わせを入力した際（左図）は，損失値が 5.2 から 1.4 へと大幅に低下しており，新たな知識が獲得されていることが分かる．一方で，編集対象ではない主語を用いた入力（右図）についても損失値が 5.1 から 2.8 に下がっており，異なる主語であっても誤って同じ目的語を出力する方向へモデルが変化していることがうかがえる．これは知識転移の陰で，関連度の近い既存知識が上書きされて

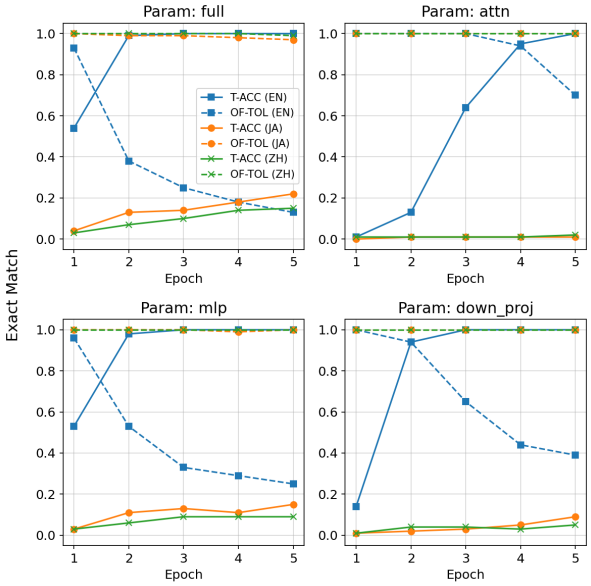

図 1　述語「首都」について異なるパラメータを更新したときのエポックごとの T-ACC と OF-TOL の推移．すべてのパラメータを更新 (full)，Attention 層のみを更新 (attn)，MLP 層のみを更新 (mlp)，MLP 層の中の down_proj のみを更新 (down_proj) の 4 パターンについてソース言語を英語として評価した．

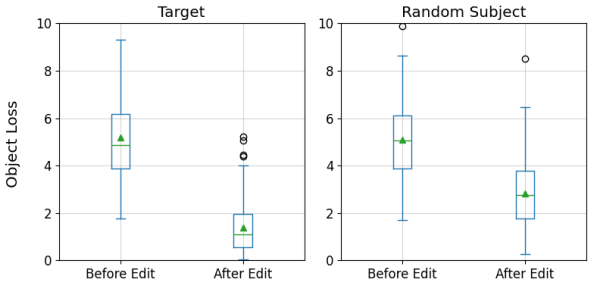

図 2　述語「首都」の 100 サンプルにおいて，英語で 5 エポックの知識編集を実施する前後で，ターゲット言語を日本語とした場合の目的語に対する損失値の変化を示す．

いることを意味する．

4　結論

本研究では，クロスリンガル知識編集手法を評価するためのデータセットを構築し，過剰適応を評価する新たな指標を設計して，LLM のクロスリンガル知識編集能力を検証した．実験の結果，編集対象の述語や更新パラメータによって転移率と副作用が大きく変動することが判明した．さらに，知識トリプルの対訳関係が明確な場合には転移効果が高まることが示唆され，LLM 学習時の言語間転移メカニズムに関する新たな示唆が得られた．今後は，言語間で知識転移しやすい条件をさらに調査し，マルチリンガル LLM のふるまいについてより深く探求したい．

謝辞

日立製作所の清水正明氏には，計算機環境の維持管理に多大なご尽力をいただき，深く感謝申し上げる．

参考文献

[1] Yuchao Li, Fuli Luo, Chuanqi Tan, Mengdi Wang, Songfang Huang, Shen Li, and Junjie Bai. Parameter-efficient sparsity for large language models fine-tuning, 2022.

[2] 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, 2024.

[3] Chenyang Song, Xu Han, Zheni Zeng, Kuai Li, Chen Chen, Zhiyuan Liu, Maosong Sun, and Tao Yang. Conpet: Continual parameter-efficient tuning for large language models, 2023.

[4] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Knowledge neurons in pretrained transformers. **CoRR**, Vol. abs/2104.08696, , 2021.

[5] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. **CoRR**, Vol. abs/2110.11309, , 2021.

[6] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

[7] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale, 2022.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[9] OpenAI. Gpt-4 technical report. **ArXiv**, Vol. abs/2303.08774, , 2023.

[10] Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. Cross-lingual knowledge editing in large language models, 2024.

[11] Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Mlake: Multilingual knowledge editing benchmark for large language models, 2024.

[12] Jiakuan Xie, Pengfei Cao, Yuheng Chen, Yubo Chen, Kang Liu, and Jun Zhao. Memla: Enhancing multilingual knowledge editing with neuron-masked low-rank adaptation, 2024.

[13] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In Roger Levy and Lucia Specia, editors, **Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)**, pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[14] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. **Commun. ACM**, Vol. 57, No. 10, pp. 78–85, September 2014.

[15] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities, 2023.

[16] Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark, 2023.

[17] Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. Easyedit: An easy-to-use knowledge editing framework for large language models, 2024.

表 2　事実に基づいた知識トリプルの例

言語	述語	主語	目的語
EN	capital	Switzerland	Bern
JA	首都	スイス	ベルン
ZH	首都	瑞士	伯尔尼
EN	architectural style	Bremen Cathedral	Gothic architecture
JA	建築様式	ブレーメン大聖堂	ゴシック建築
ZH	建築風格	不梅大教堂	哥特式建筑
EN	arterial supply	skull	external carotid artery
JA	動脈支配	頭蓋	外頸動脈
ZH	动脉供 应	颅骨	外颈动脉
EN	composer	Juditha triumphans	Antonio Vivaldi
JA	作曲者	勝利のユディータ	アントニオ・ヴィヴァルディ
ZH	作曲者	猶迪的勝利	安东尼奥·维瓦尔第
EN	director	Inception	Christopher Nolan
JA	監督	インセプション	クリストファー・ノーラン
ZH	导演	盗梦空 间	克里斯托弗·诺兰

A　知識トリプルの例

表 2 は事実に基づいた知識トリプルの例を表す．また，式 (2) における x は各言語における主語と述語を用いて以下の形式で作成する:

- 英語: The { 述語 } of { 主語 } is
- 日本語: { 主語 } の { 述語 } は
- 中国語: { 主語 } 的 { 述語 } 是

B　実験設定の詳細

知識編集時のハイパーパラメータ　クロスリンガルな知識編集に関しては未解明の点が多いため，本研究では特殊な知識編集手法を用いるのではなく，通常のモデル学習時の設定に近い条件を採用して検討を行った．

具体的な実験設定は以下の通りである:

- **Optimizer**: AdamW
- **Learning Rate**: 1×10^{-5}
- **Weight Decay**: 0.1
- **Adam Beta1**: 0.9
- **Adam Beta2**: 0.95
- **Gradient Clipping**: 1.0

また，クロスエントロピー損失の算出および勾配の計算においては，目的語トークンのみに基づいて処理を行った．

テキスト生成時のパラメータ　生成時には改行トークンまたは文末トークンが生成されるまで最大 16 トークンを貪欲法により生成した．この際，repetition penalty は設定しなかった．

表 3　Llama-3-8B における各モデルパラメータの数

パラメータの種類	パラメータの数
full	8.03B
mlp	5.64B
down_proj	1.88B
attn	1.34B

C　モデルパラメータ数

Llama-3-8B における各モデルパラメータの数を表 3 に示す．`full`（全パラメータを更新），`mlp`（MLP 層のみ），`down_proj`（MLP 層の down_proj 部のみ），`attn`（Attention 層のみ）の順にモデルパラメータが多い．