知識編集が confidence calibration へ与える影響

長谷川 遼　坂井 優介　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{hasegawa.ryo.hp5, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

言語モデルの大規模化が進むにつれ，再学習無しで知識更新が可能な知識編集の需要が高まっている．しかし知識編集は事前学習で獲得したトークン予測確率を事後学習で変化させるため，トークン予測確率と実際の精度が乖離する可能性がある．本研究ではこの問題が実際に起きているか検証するために，知識編集前後でのトークン予測確率と実際の精度の一致度を confidence calibration の観点から計算し比較した．その結果，知識編集によりモデルの confidence calibration が変化すること，特に意味理解が必要なタスクでは精度と比べトークン予測確率を相対的に低下させる傾向があることが分かった．

1　はじめに

近年の傾向として，パラメータ数が大きく大規模なデータセットで学習された大規模言語モデル (LLM) [1, 2, 3] が登場している．これらは広範な知識を持ち様々な用途に適用可能な一方，事実に基づかない情報を出力するハルシネーションの発生や，学習に長時間を要する等の課題もある．特に時間経過に伴って事実性が変わる知識に関しては，モデル全体の再学習による知識更新は困難である．

これらの課題を解決する手法の一つに知識編集 (Knowledge Editing) がある．知識編集は，言語モデル内部およびプロンプトに対し何らかの操作を加え，出力を意図通りに変化させる手法を指す．これにより，正しい事実への編集によるハルシネーションの低減や，時間経過に伴い事実性が変わる知識の獲得が容易になる．

ただし知識編集にも課題があり，特にモデルの信頼性低下がある [4]．生成系の言語モデルではプロンプトが入力された際に，出力語彙に含まれる全トークンに対して各時刻での予測確率を計算し，予測確率が最も高いトークンを出力する．このトークン予測確率と精度が一致するようなモデルは信頼性が高いと見做せる．ここで，予測確率は本来訓練データの事前学習によって獲得される．知識編集では予測確率を事後学習によって変化させるため，予測確率と精度が一致せず，モデルの信頼性を損う可能性がある．

そこで本研究では，言語モデルに対し様々な知識編集手法を適用した後に confidence calibration [5] を計算し，知識編集によるモデルの信頼性への影響について分析した．confidence calibration ではトークン予測確率と精度の一致度を計算しており，言語モデルの信頼性が実際にどの程度低下しているか検証できる．実験の結果，知識編集後は confidence calibration が変化すること，特に意味理解が必要なタスクでは，精度と比較してトークン予測確率が相対的に低くなる傾向があると明らかになった．

2　confidence calibration

信頼性の高いモデルの条件の 1 つに，予測確率 (confidence) と精度 (accuracy) の差が小さい点がある．例えば言語モデルに QA タスクを解かせる場合，トークン予測確率が 30%程度の質問が複数あるなら，それらの質問のうち正解は 30%程度であるのが望ましい．confidence calibration では上記 confidence と accuracy を比較する．confidence と accuracy の差が小さいモデルは calibration が良い，差が大きいモデルは calibration が悪いとされ，後者のうち confidence が高いモデルは overconfidence，低いモデルは underconfidence とされる．

評価には，横軸に confidence，縦軸に accuracy をとった calibration plot 図や，Expected Calibration Error (ECE) 等の §4 で紹介する指標が用いられる．

言語モデルを改善するための処理によって confidence calibration が悪化する例としては，Reinforcement Learning from Human Feedback (RLHF) [6] がある．RLHF では事後学習として，人間にとって望ましい回答の出力確率が高くなるよう強化学習が行われており，知識編集と同様に訓練データ由来の頻

度情報とは無関係にトークン予測確率を変化させている．文献 [1] では，RLHF 実施前の GPT-4 と実施済み GPT-4 の双方について，TruthfulQA の選択タスク [7] における calibration が RLHF 実施後に悪化し，overconfidence に陥ると報告している．知識編集も RLHF と同様に事前学習で獲得した予測確率を事後学習によって変化させるため，calibration へも同様に悪影響を与える可能性がある．

3 知識編集手法の分類

知識編集は主に，局所変更ベース手法，全体最適化ベース手法，外部記憶ベース手法の 3 種類に分類される [8].

3.1 局所変更ベース手法

局所変更ベース手法では，知識に強く関連するパラメータを特定し更新することで出力を変化させる．ごく一部のパラメータのみを編集するため，メモリ効率や計算効率が高い．ROME [9], MEMIT [10], KN [11] 等が該当する．

このうち ROME では，以下の 2 段階で編集が行われる．まずはモデルのフィードフォワード（FF）層の隠れ状態を分析し，出力トークンに対する寄与度を計算する．次に，寄与度が高い FF 層の重みを知識の key と value のペアに対応するメモリとみなし，新しい key と value を挿入するように編集する．

3.2 全体最適化ベース手法

全体最適化ベース手法では，知識編集時のモデル全体のパラメータ変化を学習し，他の知識の編集にも適用可能にしている．汎用性が高いが，編集パラメータ数が多く計算コストが高い．MEND [12], InstructEdit [13] などが該当する．

このうち MEND では，事前学習モデルの重みを編集するための軽量なハイパーネットワークをあらかじめ学習させ編集に使用する．学習には 1 ランクまで分解した事前学習モデルの勾配が使われる．

3.3 外部記憶ベース手法

外部記憶ベース手法では，新しい知識を外部メモリに保存し，モデルへの入力時に使用する手法である．モデルのパラメータの変更は行わない．メモリを追加すれば新たな知識編集が可能なので，拡張性に優れている．IKE [14], SERAC [15] 等が該当する．

このうち IKE では，編集したい新しい知識をプロンプトに明示的に挿入し言語モデルに入力することで，新しい知識を出力させる．

4 実験設定

実装に関する詳細な設定は §A に記載した．

4.1 データセット

本実験では，WikiDatacounterfact [16] に対して評価実験を行った．WikiDatacounterfact は WikiData を編集して作成されたデータセットで，主語-述語-目的語の triplet からなるサンプル群が含まれている．本実験では以下の 4 要素を使用した．

- 元プロンプト p: triplet のうち，主語-述語からなるプロンプト．
  (例: 'The name of the country of citizenship of Leonardo DiCaprio is')
- 言い換えプロンプト p^*: 元プロンプトを主語-述語の意味を変えずに言い換えたプロンプト．
  (例: 'Leonardo DiCaprio's country of citizenship is known as')
- 編集前正解 a_{old}: 各プロンプトの直後に来る目的語．事実と一致する．
  (例: 'United States of America')
- 編集後正解 a_{new}: 各プロンプトの直後に来る目的語．事実ではない．このエンティティをモデルが出力するように知識編集を行う．
  (例: 'Syria')

知識編集時には，モデル M_{old} が元プロンプト p に対し a_{new} を出力するよう編集し，編集後モデル M_{new} を得た．評価時には，編集後モデル M_{new} への入力を元プロンプト p および言い換えプロンプト p^* とし，正解を a_{new} として評価した．また比較のため，編集前モデル M_{old} への入力を元プロンプト p および言い換えプロンプト p^* とし，正解を a_{old} として評価した．評価指標の詳細は §4.4 で述べる．

4.2 知識編集手法

本実験では，知識編集手法のうち局所変更ベース手法として ROME，全体最適化ベース手法として MEND，外部記憶ベース手法として IKE をそれぞれ採用した．

4.3 言語モデル

本実験では，知識編集が可能なオープンソースの言語モデルとして Llama2-7B および Llama2-7B-

chat [2] を使用した．Llama2-7B-chat は Llama2-7B をベースとして，Supervised Fine-Tuning (SFT) および Reinforcement Learning with Human Feedback (RLHF) でファインチューニングされ，チャット形式で生成文を出力するように調整されている．

4.4 評価指標

本実験では，confidence calibration の評価指標として Expected Calibration Error (ECE) [5], および Adaptive Calibration Error (ACE) [17] を使用した．ECE および ACE を正解/不正解の 2 値分類タスクの評価に使用する場合は以下の式で表される．

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)| \tag{1}$$

$$ACE = \frac{1}{R} \sum_{r=1}^{R} |acc(r) - conf(r)| \tag{2}$$

ここで b は確率区間 [0, 1] を等間隔の部分区間に分割した各 bin，n_b は bin$b$ に含まれるサンプル数，B は bin$b$ の総数， r は各サンプルを予測確率でソートし等分した各 bin，R は bin$r$ の総数，N は全サンプル数，acc は bin の accuracy，$conf$ は bin 内のサンプルの confidence 即ち予測確率の平均である．両指標とも値が低いほど calibration が良いことを示す．

ECE と ACE では bin の分割方法が異なる．各 bin に属するサンプル数が ECE では異なる一方 ACE では等しいため，生成確率の偏りに対して ACE の方がよりロバストな指標と言える．

また calibration の傾向が overconfidence と underconfidence のいずれかを評価する指標として，(1) 式で絶対値の代わりに $conf(b) - acc(b)$ を計算する Miscalibration Score (MCS) [18] がある．MCS が高いほど overconfidence が優勢であることを示す．

$$MCE = \sum_{b=1}^{B} \frac{n_b}{N} (conf(b) - acc(b)) \tag{3}$$

本実験では bin 数を 10 として評価した．

5 結果と考察

5.1 Accuracy および confidence

表 1，2 に知識編集時の Accuracy，ECE，ACE の値を，モデル，編集手法，元プロンプト/言い換えプロンプトに分けて示す．

Accuracy について，ROME および IKE ではモデルやプロンプトの種類に関わらず知識編集前と比較して上昇している一方，MEND では知識編集前より低下している．このことから，ROME と IKE では知識の編集に成功し，MEND では今回の実験ではうまく知識が編集できていない可能性が高い．

表 1 元プロンプト p 使用時の accuracy/ECE/ACE

	Llama-2-7b			Llama-2-7b-chat		
	accuracy	ECE	ACE	accuracy	ECE	ACE
知識編集前	0.323	0.141	0.149	0.224	0.116	0.116
ROME	0.936	0.026	0.018	0.928	0.031	0.028
MEND	0.287	0.035	0.024	0.409	0.031	0.039
IKE	**1.000**	**0.003**	**0.003**	**0.994**	**0.018**	**0.017**

表 2 言い換えプロンプト p^* 使用時の accuracy/ECE/ACE

	Llama-2-7b			Llama-2-7b-chat		
	accuracy	ECE	ACE	accuracy	ECE	ACE
知識編集前	0.185	**0.052**	**0.043**	0.217	0.104	0.104
ROME	0.528	0.104	0.105	0.403	**0.056**	0.059
MEND	0.081	0.072	0.072	0.137	0.209	0.209
IKE	**0.846**	0.076	0.073	**0.785**	0.057	**0.058**

また知識編集後の accuracy に関して，言い換えプロンプトよりも元プロンプトの方が高い．知識編集時と評価時のプロンプトが同一の場合，モデルはプロンプトの意味だけでなく表層形から出力トークンを判断できる．入力の表層形を使用できる問題設定では意味のみを使用する場合よりも難易度が明確に低いことが示された．

ECE について，元プロンプトでは，Llama-2-7b と Llama-2-7b-chat の双方で知識編集前と比較するとどの編集手法でも良化している．知識編集時と評価時のプロンプトが同一の場合は accuracy のみならず calibration でも良い結果が得られている．

一方で，言い換えプロンプトの ECE に関しては以下のような結果となった．Llama-2-7b ではどの編集手法でも編集前と比較して悪化している．一方 Llama-2-7b-chat では ROME と IKE で良化し，MEND で悪化している．知識編集時と評価時のプロンプトの表層形が異なり，意味理解が必要になる場合は，必ずしも calibration が良化するとは限らず，モデルによって傾向が異なることが分かった．

ACE については，編集手法・モデル・元プロンプト/言い換え文問わず ECE と同じ傾向が見られた．

以上より，知識編集によって accuracy は上昇するものの，calibration は言い換えなどの表層形の揺らぎに対して頑健でなく，意味理解が必要な場合は悪化する可能性があると示された．また言い換えプロンプト使用時の calibration の良化/悪化は手法ではなくモデルに依存することが分かった．

表 3　言い換えプロンプト使用時の MCE

	Llama-2-7b	Llama-2-7b-chat
知識編集前	-0.032	0.104
ROME	-0.104	0.031
MEND	0.072	0.209
IKE	-0.073	0.010

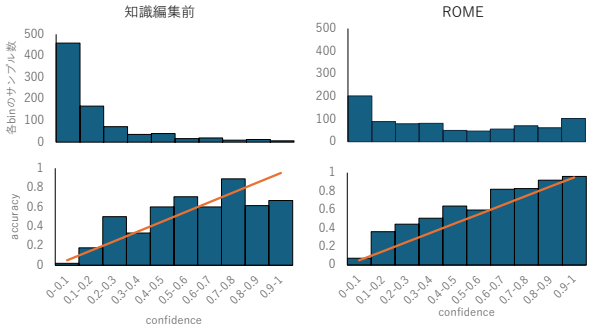

図 1　Llama-2-7b における知識編集前及び ROME の calibration plot，上段は各 bin のサンプル数

5.2　Over/underconfidence の傾向

§5.1 でモデルごとに傾向が異なった言い換えプロンプトについて詳細に分析するため，表 3 に MCS を示す．知識編集前では，Llama-2-7b で calibration が良く，Llama-2-7b-chat で overconfidence の傾向が見られた．知識編集後は ROME と IKE では Llama-2-7b で underconfidence，Llama-2-7b-chat で calibration が良い傾向が見られた．一方で知識が獲得できていない可能性があると示された MEND では，双方とも overconfidence の傾向が見られた．

知識編集前及び ROME の calibration plot について，図 1 に Llama-2-7b の場合を，図 2 に Llama-2-7b-chat の場合を示す．知識編集前では confidence の低い bin にサンプルが偏り，その bin の calibration が全体の calibration に大きく影響していた．特に Llama-2-7b では全体の 5 割以上に当たる 459 サンプルが confidence の最も低い bin に属していた．

一方で，ROME による知識編集後は confidence が高い bin のサンプル数が増えた．また知識編集前と比較して各 bin の accuracy が全体的に上昇した．そのため calibration が良かった Llama-2-7b では underconfidence に，overconfidence だった Llama-2-7b-chat で calibration が良くなった．IKE でも同様の傾向が見られた (付録 B)．

以上から，知識編集によって confidence 即ちトークン予測確率は上昇するものの，accuracy の上昇幅よりは小さいため，underconfidence 方向の変化をもたらすと分かる．知識編集によるモデルの知識獲得はトークン予測確率には反映されづらいと言える．

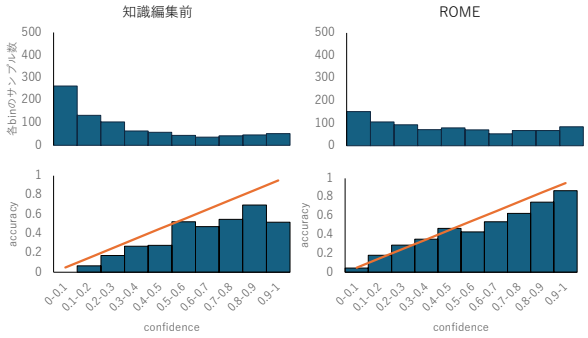

図 2　Llama-2-7b-chat における知識編集前及び ROME の calibration plot，上段は各 bin のサンプル数

最後に，知識編集と RLHF との関連について述べる．前述の通り知識編集前の Llama2-7b-chat は overconfidence で，これは §2 の GPT-4 の例と一致する．知識編集と RLHF は事前学習で獲得した予測確率を事後学習によって変化させ，その結果 calibration に影響を及ぼす点で共通している一方，知識編集は underconfidence，RLHF は overconfidence を引き起こす点で異なることが分かった．知識編集後の Llama2-7b-chat では両効果が相殺され，結果的に calibration が良化したと考えられる．

6　おわりに

本研究では，言語モデルの confidence calibration に対して知識編集が与える影響について分析した．その結果以下の知見が得られた．

- 編集時と評価時のプロンプトが同一の場合は知識編集により calibration が良化するが，言い換えにより表層形が異なり意味理解が必要になる場合は悪化することがある．
- 意味理解が必要な場合は，精度と比べトークン予測確率を相対的に低下させる underconfidence 方向の変化をもたらす．知識編集による知識の獲得はトークン予測確率には反映されづらいことを意味する．

今後の方針としては追加実験がある．Llama-2-7b および Llama-2-7b 以外のモデル，ROME/IKE/MEND 以外の知識編集手法，WikiDatacounterfact 以外のデータセットで追加実験を行い，本研究と一貫性を持った傾向が観測されるか検証したい．

謝辞

本研究は JSPS 科研費 JP23H03458 の助成を受けたものです.

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. **arXiv preprint arXiv:2309.16609**, 2023.

[4] Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in llms? **arXiv preprint arXiv:2406.19354**, 2024.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In **International conference on machine learning**, pp. 1321–1330. PMLR, 2017.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. **Advances in neural information processing systems**, Vol. 30, , 2017.

[7] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. **arXiv preprint arXiv:2109.07958**, 2021.

[8] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. **ACM Computing Surveys**, Vol. 57, No. 3, pp. 1–37, 2024.

[9] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 17359–17372, 2022.

[10] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. **arXiv preprint arXiv:2210.07229**, 2022.

[11] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. **arXiv preprint arXiv:2104.08696**, 2021.

[12] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. **arXiv preprint arXiv:2110.11309**, 2021.

[13] Ningyu Zhang, Bozhong Tian, Siyuan Cheng, Xiaozhuan Liang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, and Huajun Chen. Instructedit: Instruction-based knowledge editing for large language models. **arXiv preprint arXiv:2402.16123**, 2024.

[14] Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? **arXiv preprint arXiv:2305.12740**, 2023.

[15] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In **International Conference on Machine Learning**, pp. 15817–15831. PMLR, 2022.

[16] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. **arXiv preprint arXiv:2401.01286**, 2024.

[17] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In **CVPR workshops**, Vol. 2, 2019.

[18] Shuang Ao, Stefan Rueger, and Advaith Siddharthan. Two sides of miscalibration: identifying over and under-confidence prediction for network calibration. In **Uncertainty in Artificial Intelligence**, pp. 77–87. PMLR, 2023.

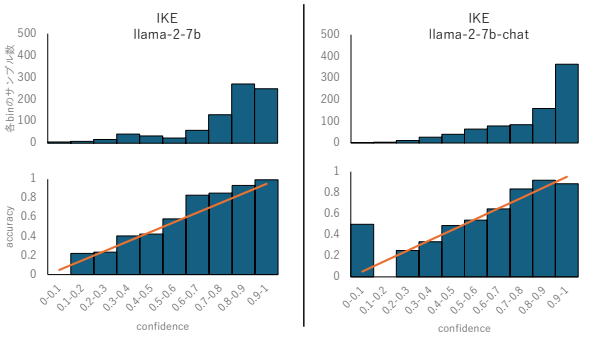

図 3 Llama-2-7b と Llama-2-7b-chat における IKE の calibration plot，上段は各 bin のサンプル数

A 実験の詳細設定

知識編集時および評価時には，知識編集フレームワークとして EasyEdit [16] を使用した．GPU は編集前モデルの評価時および ROME での知識編集時と評価時には GeForce RTX 3090 を，MEND および IKE での知識編集時と評価時には NVIDIA A100 80GB PCIe をそれぞれ 1 台使用した．また実装には HuggingFace を使用し，言語モデルは Llama2-7b が meta-llama/Llama-2-7b-hf，Llama2-7b が meta-llama/Llama-2-7b-chat-hf を用いた．

B IKE の over/underconfidence の傾向

IKE による編集後の calibration plot について，図 3 に示す．§5.2 で言及した通り，confidence が高い bin のサンプル数が増え，知識編集前と比較して各 bin の accuracy が全体的に高いという ROME と同様の傾向が見られた．