言語モデルにおける知識の連鎖的編集を促すメタ学習手法

芝太地 [1,2]　磯沼大 [2,3]　大関洋平 [1,2]　宮尾祐介 [1,2]
[1] 東京大学 [2]NII LLMC [3] 東北大学
{shiba-taichi60,yusuke}@is.s.u-tokyo.ac.jp
oseki@g.ecc.u-tokyo.ac.jp　isonuma@nii.ac.jp

概要

大規模言語モデルの知識編集は，現実世界で頻繁に変化する事実をモデルに反映させる上で重要な課題である．しかし，既存の知識編集手法は，編集対象の知識を修正できる一方で，その修正が論理的に関連する知識へ十分に伝播しない．本研究では，関連知識への伝播を実現する知識編集アルゴリズムを設計するのではなく，知識編集を適用するモデルに着目し，「新知識の学習が関連知識の学習を伴う」ような内部表現を獲得することを目的とするメタ学習手法を提案する．実験の結果，提案手法により得られたモデルは，新知識のみを用いた単純な勾配降下による学習であっても，関連知識まで編集が伝播しやすいことが示された．

1　はじめに

大規模言語モデルは，大規模なテキストコーパスを用いた事前学習を通して，多様な世界知識を保持している．しかし，モデルが保持している知識は事前学習時点に依存した静的な情報であり，現実世界で頻繁に変化する知識を反映することはできない．この課題に対し，近年モデルが保持する知識を直接編集する知識編集手法が数多く提案されている [1, 2, 3]．しかし，知識編集における重要な課題として，編集対象の知識は効率的に追加・修正できるものの，対象知識と論理的に関連する知識までは十分に編集が反映されない点が指摘されている [4, 5]．例えば，ある人物の出身国に関する知識を編集した場合でも，その人物の出身国の通貨や首都に関する新知識を用いたマルチホップ推論を要する知識まで編集が十分に伝播しない．このような問題は知識の一貫性や推論能力の低下につながる．

この課題を解決するため，既存研究の多くは，新しい知識が与えられた際にどのように関連知識へと情報を伝播させるかに焦点を当ててきた [6, 7]．しかし，これらの手法はコストが大きいことや，学習時のデータセットに依存しやすく汎化性が低いことなどの問題を抱えている．

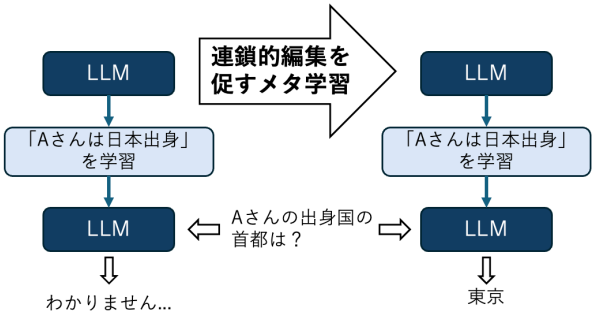

図 1　本研究では新知識の学習が関連知識まで伝播する内部表現を獲得するメタ学習手法を提案する．

本研究では，新しい知識を「どのように関連知識へ伝播させるか」という編集アルゴリズムの設計ではなく，知識編集を適用するモデルに着目する．すなわち，図 1 に示すように「新たな知識を学習した際に，関連知識まで自動的に伝播して学習するようなパラメータをモデルが持ちうるのではないか」という仮説に基づき，モデルパラメータを事前にチューニングするアプローチを提案する．具体的には，Model-Agnostic Meta-Learning（MAML）[8] に基づくメタ学習により，知識編集に対して適応的なモデルを獲得する．

実験では，2 ホップ推論を要する質問（第 1 ホップ：新しい知識，第 2 ホップ：一般常識）からなるデータセット [6] を用い，事前学習済みモデルのメタ学習を行う．得られたモデルに対し，新しい知識のみを用いた単純な勾配降下による学習が，どの程度関連知識まで伝播するかを検証する．その結果，少なくとも訓練データと同一の推論形式（第 1 ホップ：新しい知識，第 2 ホップ：一般常識）において，提案手法は追加的な編集機構を用いずとも，編集された知識が論理的に関連する知識へ伝播しやすい内部表現を誘導できることが確認され，伝播性の向上

と学習効率を両立する新たな知識編集の枠組みとして有効であることが示された．

2 関連研究

新しい知識をモデル内部のパラメータに反映する知識編集手法として，ROME [1]，MEMIT [2]，AlphaEdit [3] などが提案されている．これらは編集対象の知識に対して局所的にパラメータを修正することで効率的な編集を実現する一方，編集に伴って整合性を保つべき関連知識にまで編集が十分に伝播しないことが指摘されている [4, 5]．

関連知識への伝播性を高める試みとして，CaKE [7] は，新知識の言い換え表現や新知識を用いたマルチホップ推論を通じて学習データを増強し，編集後の知識の一貫性向上を図る．しかし，このアプローチは新知識ごとに追加データを生成・収集する必要があり，運用上のコストが大きい．一方，PropMEND [6] は，新知識に基づく勾配を関連知識の編集につながる勾配へ写像するハイパーネットワークを学習することで，編集の伝播を促進する．しかし，ハイパーネットワークの学習時に観測されないエンティティや関係に対しては，関連知識への伝播が限定的となり，汎化性が課題として残る．本研究は，新たな知識を学習する際の追加データ生成や外部機構に依存するのではなく，モデルを事前に「関連知識へ伝播しやすい」パラメータへ調整する観点から，知識の伝播性を促す手法を提案する．

3 提案手法

本研究では，事前学習済みモデルを「関連知識まで編集が伝播しやすいモデル」に調整する．

モデル f_θ において，編集対象の新しい知識を表す $target$ と，編集時に同時に整合性を保つべき関連知識 $related$ からなる知識編集タスク $\tau = (target, related)$ を考える．理想的には，任意のタスクについて，$target$ を学習した際に，$related$ に関する振る舞いも一貫して編集されるような初期パラメータ θ_{meta} を獲得したい．これを次式で定式化する：

$$\theta_{meta} = \arg\min_{\theta} \mathbb{E}_{\tau=(target,related)\sim p(\tau)} \left[\mathcal{L}_{related}\left(f_{\theta_\tau}\right) \right] \tag{1}$$

ただし，

$$\theta_\tau = \theta - \alpha \nabla_\theta \mathcal{L}_{target}\left(f_\theta\right). \tag{2}$$

ここで f_{θ_τ} は，$target$ に対する損失 $\mathcal{L}_{target}$ を用いて勾配降下法を数ステップ適用して得られる新知識学習後のモデルを表す．すなわち式 (1) は，「新知識 $target$ を学習することで，関連知識 $related$ も整合的に学習されるパラメータ」を得ることを意味する．

この最適化は MAML [8] に基づき，以下の 2 段階ループで実行する：

1. タスク分布 $p(\tau)$ から知識編集タスク $\tau = (target, related)$ をサンプリングする．
2. 内側ループ：$target$ のみを用いてパラメータを更新し，編集後パラメータ θ_τ を得る（式 (2)）．
3. 外側ループ：新知識学習後のモデル f_{θ_τ} に対する $related$ の損失が小さくなるように，初期パラメータ θ を更新する：

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{related}\left(f_{\theta_\tau}\right). \tag{3}$$

4. 上記を反復する．

通常の MAML では，式 (3) の勾配計算に θ に関する二階微分が含まれるため，LLM のような大規模モデルでは計算コストとメモリ負荷が大きい．そこで本研究では，二階微分項を無視する First-Order MAML（FOMAML）[8] による近似

$$\nabla_\theta \mathcal{L}_{related}\left(f_{\theta_\tau}\right) \approx \nabla_{\theta_\tau} \mathcal{L}_{related}\left(f_{\theta_\tau}\right) \tag{4}$$

を採用し，実用的な計算量でメタ学習を実現する．

4 評価実験

本節では，提案するメタ学習手法により得られたモデルが，新しい知識を与えられた際に，編集対象の知識だけでなく論理的に関連する知識へも編集が伝播するような性質を獲得できているかを検証する．

実験設定　メタ学習における内側ループおよび外側ループの最適化には，いずれも AdamW [9] を用いた．学習率は内側ループを $\alpha = 2 \times 10^{-5}$，外側ループを $\beta = 1 \times 10^{-5}$ とし，内側ループの更新ステップ数は 10，バッチサイズは 1 とした．また，外側ループは 1 エポック実施，すなわち，訓練データに含まれるタスクをそれぞれ 1 回ずつサンプリングして，各タスクにつき内側ループ更新と外側ループ更新を行った．なお，内側ループおよび外側ループの学習における損失計算では，先行研究 [10] に倣い，対象知識の全トークンに対して損失を計算するのではなく，目的語（質問応答では正答）に対応するトークン列にのみクロスエントロピー損失を課した．

データセット　Controlled Ripple Edit（CRE）[6] を用いた．CRE は，架空のエンティティに関する新しい事実と，その新知識を 1 ホップ目，モデルが事前に保持していると想定される知識を 2 ホップ目とした 2 ホップ推論を要する質問から構成される知識編集データセットである．本研究では，新しい知識を学習させたモデルが 2 ホップの質問に正しく応答できるかを評価した．

CRE には，訓練データ 4,000 件，分布内評価データ 500 件，分布外評価データ 350 件が含まれる．分布外評価データは，訓練データに出現しないエンティティやリレーションを含み，正答には編集した新知識とモデルの既存知識の双方を適切に組み合わせた推論が必要となる．具体的なデータセットの例は付録 A に示す．提案手法のメタ学習には訓練データを用い，メタ学習後のモデルの評価には分布内・分布外評価データを用いた．

さらに，より広範なマルチホップ推論を含む知識編集データセットである MQuAKE [5](3,000 件) も追加の評価データとして用いた．

評価指標　先行研究 [7] に従い，評価指標には正解率を用いた．各評価質問に対して，新知識学習後のモデルで最大 10 トークンを生成し，期待される正答が出力に含まれる場合を正解と判定した．

4.1　学習対象による違い

提案手法はメタ学習により「関連知識への伝播」を促すことを目的とするが，その効果は更新対象とするパラメータに依存する可能性がある．そこで本節では，どの層（パラメータ集合）を学習対象とすることが，分布外条件における汎化性能の向上に有効かを検証する．

設定　事前学習済みモデルとして Llama3-8B-Instruct [11] を用いた．更新範囲として，全層更新と単層更新を比較する．さらに，各更新範囲について，更新対象モジュールを Attention のみ，MLP のみ，Attention+MLP の 3 条件で切り替えた．単層更新では，指定した層以外のパラメータは固定し，当該部分のみを学習対象とした．

結果　まず，全層更新の結果を表 1 に示す．MLP を更新対象に含む条件では，分布内評価に対して正解率が向上する傾向が観察された．一方，分布外評価における改善は限定的であった．これは，全層更新では更新が広範囲に及び，訓練データに過度に適合した結果，未出現エンティティや未出現リレーションを含む分布外条件で汎化しにくくなった可能性を示唆する．

表 1　全層を学習対象とした CRE の分布内・分布外評価における正解率．ただし，括弧内はメタ学習による正解率の上昇を示す．

	分布内評価（%）	分布外評価（%）
Attn	18.6 (+11.8)	2.1 (+1.6)
MLP	71.6 (+65.9)	6.3 (+6.3)
Attn + MLP	52.5 (+46.6)	5.7 (+5.7)

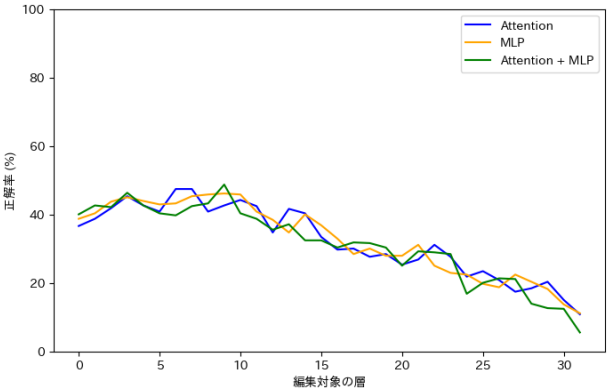

図 2　単層更新による CRE の分布外評価の結果．

次に，単層更新の分布外評価結果を図 2 に示す．その結果，更新対象モジュール（Attention，MLP，Attention+MLP）に依らず，浅い層を学習対象とすると分布外評価に対する正解率が高くなる傾向が確認された．単層更新は更新範囲が局所的であるため，モデルの既存知識構造を大きく破壊せずに編集を適用でき，結果として既存知識と新しい知識を組み合わせるマルチホップ推論能力の向上に寄与したと考えられる．

以上より，以降の実験では，本節で最も良好な性能を示した浅い層（Llama3-8B では第 9 層）に対する単層更新（Attention+MLP）を基本設定として採用し，提案手法および比較手法の評価を行う．

4.2　既存手法との比較

本節では，提案手法が既存の知識編集手法と比較して，関連知識への伝播に有効であるかを検証する．比較対象として，代表的なパラメータ編集手法である ROME [1]，MEMIT [2]，AlphaEdit [3], CaKE [7]，および，事前学習済みモデルにメタ学習を適用せずに提案手法における Inner Loop のみを適用した Fine-Tuning (FT) を用いる．すべての手法は同一の事前学習済みモデル Llama3-8B 上で評価し，比較手法は EasyEdit [10] の公開実装を用い，ハイパーパ

表 2　既存手法と提案手法を複数のデータセットで比較した結果．時間は MQuAKE-subset の各編集に要する実行時間を示す．

	CRE 分布外 (%)	MQuAKE (%)	MQuAKE-subset (%)	時間 (秒)
ROME	12.9	16.2	50.2	9.13
MEMIT	1.7	7.7	9.4	9.51
AlphaEdit	5.1	9.7	37.6	16.2
CaKE	-	52.4	23.9	45.8
FT	17.4	28.6	36.2	1.61
提案手法	47.1	49.2	73.2	1.68

ラメータは本実装の推奨設定に従った．

評価設定　評価は (i) Controlled Ripple Edit（CRE）の分布外評価データ，(ii) より広範なマルチホップ推論を含む MQuAKE 全体，(iii) MQuAKE のうち CRE と同じ推論設定，すなわち，2 ステップ推論のうち，1 ステップ目に対応する知識を編集する条件に限定したサブセット 213 件（以下，MQuAKE-subset）で行った．なお，提案手法のメタ学習に用いるデータセットについて，MQuAKE は評価用のデータセットであるため，すべての実験設定において CRE の訓練データを用いて行った．

結果　表 2 に各評価設定における正解率の結果を示す．CRE の分布外評価において，提案手法は既存手法を上回る性能を示した．一方で，MQuAKE 全体に対する評価では，CaKE と比較して提案手法の性能は低かった．これは，提案手法のメタ学習に用いた CRE が「2 ホップ推論の 1 ホップ目の知識を編集する」という特定の形式に限定されており，MQuAKE が含む多様な推論形式を十分にカバーできていないことが一因である可能性がある．実際に，MQuAKE を CRE と同様の推論形式に限定した MQuAKE-subset では，提案手法が既存手法を上回る結果が得られた．今後の課題として，より多様な推論形式を含むデータセットを用いてメタ学習を行った場合に，提案手法が MQuAKE のような多様な推論形式へ汎化できるかを検証する必要がある．

さらに，1 編集あたりの実行時間を比較した結果，提案手法は高い精度を達成しつつ，既存手法より短い実行時間で編集を行えることが確認された．これは，提案手法が編集時にデータ拡張や外部構造を必要とせず，学習済みの「伝播しやすい」パラメータ構造に基づいて少ない更新で実現できるためである．

以上より，提案手法は，CRE と同様の条件において，既存のパラメータ編集手法よりも高い伝播性と効率性を両立できることが示された．

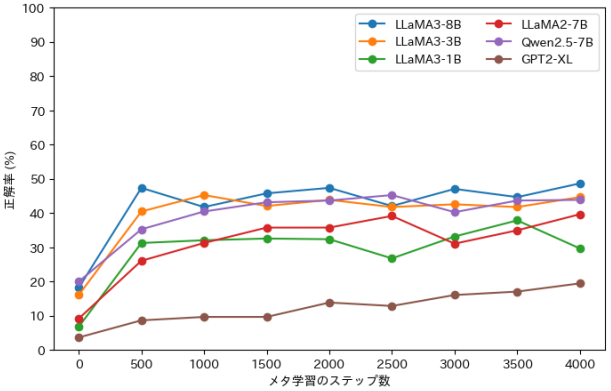

図 3　複数モデルにおける一般性．

4.3　複数モデルに対する汎用性

最後に，提案手法の効果が異なるモデル系列においても観察されるかを検証する．対象モデルとして，Llama3 (8B, 3B, 1B) [11]，Llama2-7B [12]，Qwen2.5-7B [13]，GPT2-XL [14] を用いた．各モデルの学習対象層は付録 B に示す．評価は，CRE の分布外評価データに対する正解率を対象とする．

結果　メタ学習のステップ数に対する正解率の推移を図 3 に示す．提案手法の効果はモデルによって異なり，モデルサイズが大きいほど分布外評価における正解率が高い傾向が確認された．また，同程度の規模であっても，より新しいアーキテクチャ（例：Llama3 系列や Qwen2.5）が，旧世代のモデル（例：Llama2，GPT2-XL）に比べて分布外での汎化性能が高い傾向を示した．

これらの結果は，提案手法による「関連知識への伝播性」は，モデルがもつ表現能力や知識表現の構造に影響を受けることを示唆する．

5　おわりに

本研究では，大規模言語モデルにおける知識編集の課題として，編集対象の知識は修正できる一方で，論理的に関連する知識へ編集が伝播しにくいという課題に着目した．この課題に対し，知識編集アルゴリズムの設計ではなくモデルを事前に調整する観点から，MAML に基づくメタ学習により「関連知識へ伝播しやすいパラメータ」を獲得する手法を提案した．実験の結果，提案手法により，新知識を学習した際の知識の伝播性が向上するモデルを獲得できることが示された．今後は，メタ学習の過程で知識の保持構造や推論経路がどのように変化したのかについて，詳細な分析を進める．

謝辞

本研究は，JST BOOST JPMJBY24A6，JPMJBY24B2 の支援を受けたものです．

参考文献

[1] 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.

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

[3] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained model editing for language models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[4] Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 283–298, 2024.

[5] Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 15686–15702, Singapore, December 2023. Association for Computational Linguistics.

[6] Zeyu Leo Liu, Greg Durrett, and Eunsol Choi. Propmend: Hypernetworks for knowledge propagation in llms. **arXiv preprint arXiv:2506.08920**, 2025.

[7] Yunzhi Yao, Jizhan Fang, Jia-Chen Gu, Ningyu Zhang, Shumin Deng, Huajun Chen, and Nanyun Peng. CaKE: Circuit-aware editing enables generalizable knowledge learners. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 11377–11393, Suzhou, China, November 2025. Association for Computational Linguistics.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In **International conference on machine learning**, pp. 1126–1135. PMLR, 2017.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

[10] 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.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv e-prints**, pp. arXiv–2407, 2024.

[12] 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.

[13] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

表 3　CRE における学習・評価データの例.

データ分割	編集対象知識 (*target*)	関連質問 (*related*)
訓練データ	Smith Innovation PLC's global headquarters was located in Greece.	What is the currency of the country that hosted Smith Innovation PLC's global headquarters?
分布内評価	Orange Development Inc. was founded in Greece.	What is the currency of the country that Orange Development Inc. was founded in?
分布外評価	Allen Manufacturing LLC was founded in Portugal.	Which religion has the most followers in the country that Allen Manufacturing LLC was founded in?

表 4　MQuAKE および MQuAKE-subset の例.

	編集対象知識（*taret*）	関連質問（*related*）
MQuAKE	Basketball was created in the country of (United States of America → Spain). The official language of Spain is (Spanish → Arabic).	What is the official language of the country of origin of Hapoel Eilat B.C.'s sport?
MQuAKE-subset	Joe Biden is a citizen of (United States of America → Vietnam).	Which language is the official language of Joe Biden's country?

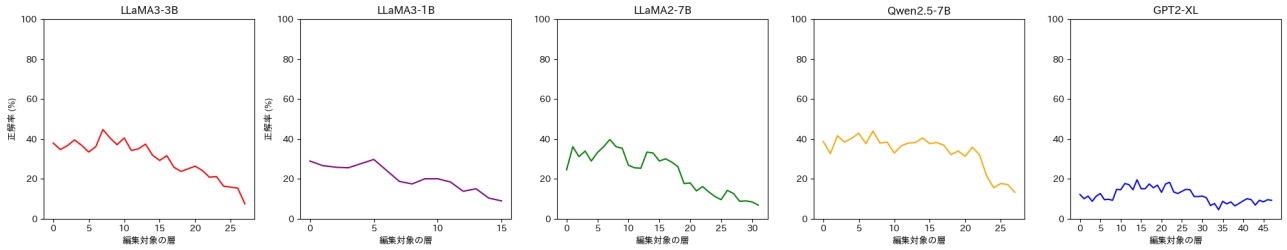

図 4　複数モデルにおける学習対象の層と分布外評価の関係.

A　データセットの具体例

本研究では，Controlled Ripple Edit（CRE）[6] を用いてメタ学習および評価を行った．表 3 に CRE の具体例を示す．分布内評価データは訓練データと同種のエンティティ（例：Greece）・リレーション（例：currency）を含むのに対し，分布外評価データは訓練データに出現しないエンティティ（例：Portugal）や関係（例：religion）を含む．

さらに，より広範なマルチホップ推論を含む知識編集データセットとして MQuAKE [5] を評価に用いた．MQuAKE は 2〜4 ホップの推論を要する質問から構成され，複数ホップにまたがる複数の事実の同時編集を含む点が特徴である．表 4 に示す例では，3 ホップ推論を要する質問に対して，第 2・第 3 ホップに対応する 2 つの事実を同時に編集している．一方，MQuAKE-subset は，2 ホップ推論において第 1 ホップ目の事実のみを編集する設定に限定した部分集合であり，メタ学習に利用した CRE と同等の推論形式から構成される．

B　複数モデルにおける学習対象とする層の選定

複数モデルにおける汎化性の評価では，Llama3-8B 以外に Llama3-3B，Llama3-1B，Llama2-7B，Qwen2.5-7B，GPT2-XL を用いた．Llama3-8B では第 9 層の Attention+MLP を学習対象としたが，他のモデルでも同様に，どの層の Attention+MLP を学習対象とすることが分布外評価で有効かを事前に検証した．その結果を図 4 に示す．図 4 に基づき，各モデルで分布外評価の正解率が最も高い層を更新対象として採用した．選定された層は，Llama3-3B は第 7 層，Llama3-1B は第 5 層，Llama2-7B は第 7 層，Qwen2.5-7B は第 7 層，GPT2-XL は第 14 層である．