排他的逆学習

佐々木睦史 [1,2]　中山功太 [2]　宮尾祐介 [3,2]　大関洋平 [3,2]　磯沼大 [2,1,3]
[1] 東北大学　[2] NII LLMC　[3] 東京大学
mutsumi.sasaki@dc.tohoku.ac.jp　{nakayama, isonuma}@nii.ac.jp
yusuke@is.s.u-tokyo.ac.jp　oseki@g.ecc.u-tokyo.ac.jp

概要

大規模言語モデルを医療や教育などに導入する際，有害な文章が生成されるリスクが課題となる．既存の逆学習では特定の有害な知識・表現を忘却できるものの，有害な知識・表現は多様であり，網羅的な除去は困難である．本研究では，忘却対象を個別に列挙する代わりに，「保持したい知識・表現」以外を広く忘却させることで広範な有害性の除去を目指す排他的逆学習を提案する．排他的逆学習により，医療や数学など特定の知識に関する多様な指示への応答能力を維持しつつ，Jailbreak を含む幅広い入力に対する安全性が担保されたモデルが得られることを示す．

1　はじめに

大規模言語モデル（LLM）は，プライバシー情報の漏洩，著作権侵害といったリスクを内包している．LLM は膨大なコーパスで学習されるため，学習データから有害な内容を事前にすべて除去することは現実的でない．このため，学習済み LLM から特定の知識を選択的に削除する逆学習 [1–5] が低コストな対策として注目されてきた．

しかし，従来の逆学習研究は「何を忘却するか」という問題設定に基づいており，忘却するべき有害な表現を網羅的に列挙する必要がある．実際にはあらゆる攻撃パターンを想定して多様な忘却データを用意することは困難であり，学習時に含まれない未見の有害質問への汎化は依然として不十分である [3, 6]．特に，近年では Jailbreak 攻撃の手法 [7–13] が多様化しており，忘却データの指定により有害な生成を促すあらゆる入力に対して防御することをますます難しくしている．

本研究では，逆学習の問題設定を転換することで，この問題の解決を試みる．すなわち，「何を忘却するか」を指定する代わりに，「何を保持するか」のみを指定し，それ以外のすべてを忘却する排他的逆学習（Exclusive Unlearning; EU）に取り組む（図 1）．具体的なアプローチとして，モデル自身から多様な文章をサンプリングし，それらの生成確率を一様分布に近づけると同時に，保持対象データセットに対する標準的なファインチューニングを行う方法を提案する．これにより，モデルが保持する知識・能力全体を忘却させつつ，対象能力を保持することが可能になる．従来の逆学習は忘却データを必要とする一方，排他的逆学習は忘却すべき有害な入力を逐一列挙する必要がないため，忘却データ依存という従来の逆学習の限界を原理的に回避できる．この問題設定は，医療診断支援や教育支援など，保持すべき能力が明確に定まるドメイン特化型 LLM の応用シナリオにおいて適用可能である．例えば医療応用では，「医療知識と医療質問に関する指示追従能力のみを保持し，それ以外は必要ない」といったように，保持対象が明確に定義できる．

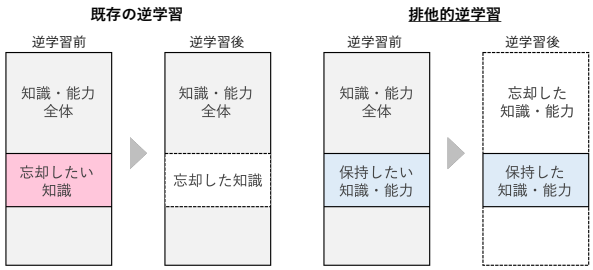

図 1　排他的逆学習の概要．「何を保持するか」のみを指定し，それ以外を排他的に忘却するため，忘却データに含まれない未見の有害領域への汎化が難しいという従来の逆学習の限界を回避できる．

実験では，医療，教育分野への応用を見据え，医療および数学の Instruction Tuning データセットを用いた実験により，本手法の有効性を実証した．排他的逆学習後のモデルは，保持対象能力を元のモデルと同等以上に維持しつつ，Jailbreak を含む幅広い有害入力に対して高い防御性能を示した．特に，防御性能において，未見の Jailbreak にも有効性を示して

きた既存の逆学習手法 [6, 14] を大きく上回る結果が得られた．さらに，本手法の設計選択の妥当性を検証し，事前学習済みモデルからの忘却が保持対象における言語生成能力の保持に必要なこと，モデルが自己生成した文章に一様損失を課す設計が広範な知識の忘却と学習の安定化に寄与することを示した．

2 提案手法

我々の目的は，データセットによって与えられる特定タスクを解く能力だけを保持し，それ以外の知識や能力はすべて忘却することである．

逆学習では基本的に忘却損失 $\mathcal{L}_{\text{forget}}(\theta)$ と保持損失 $\mathcal{L}_{\text{retain}}(\theta)$ を組み合わせたパラメータ最適化を行う．従来の逆学習の最も基本的な方法である勾配上昇法（Gradient Ascent; GA）では，事前に定義された忘却データ $\mathcal{D}_{\text{forget}}$ の対数尤度を最小化するように $\mathcal{L}_{\text{forget}}(\theta)$ を設定する．

$$\mathcal{L}_{\text{forget}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{forget}}}\left[\log p_\theta(x)\right] \tag{1}$$

しかし，事前に定義された $\mathcal{D}_{\text{forget}}$ で逆学習をしてしまうと，$\mathcal{D}_{\text{forget}}$ 以外も含むあらゆる有害入力への汎化が難しい．

そこで，本研究では，保持対象のデータセットのみを指定し，それ以外の知識・能力を排他的に忘却することで，忘却データ不要の逆学習を実現する．我々の手法の特徴は，モデル自身が出力した文章を忘却することである．モデルの自己生成文章を用いる設計は，「モデルが自己生成する出力の中に，内部の知識や能力が含まれている」という直感に基づく．忘却中は正常な文章生成能力が失われていくことから，学習前のモデル θ_0 自身からサンプルした文章を忘却する．この手法により，既存の逆学習の問題となっていた，あらゆる有害質問，攻撃パターンを想定して忘却データを作成することが困難であるという問題を解決する．一方で，保持対象以外の知識・能力は有害かどうかに関わらず忘却することになるため，本手法は医療，教育応用など保持すべき対象が明確である場合に有用となる．具体的な損失として，一様分布 $p_u(\cdot)$ と，モデルの生成確率 $p_\theta(\cdot \mid x_{<t})$ の間のクロスエントロピー（CE）として $\mathcal{L}_{\text{forget}}(\theta)$ を採用する．

$$\mathcal{L}_{\text{forget}}(\theta) = \mathbb{E}_{x\sim p_{\theta_0}}\left[\frac{1}{T}\sum_{t=1}^{T}\text{CE}[p_u(\cdot), p_\theta(\cdot|x_{<t})]\right] \tag{2}$$

ここで T はテキスト x の系列長を表し，$x_{<t}$ は時刻 t より前に生成されたトークン列を表す．この損失の考え方自体は既存研究 [15] でも採用されており，忘却時に損失が有限値に収束するため，破局的崩壊を回避できる．実際，§ 5 で，既存の勾配上昇法による損失ではなく，一様損失を用いる妥当性を示す．

保持項については，指定した能力を保持するために，標準的なファインチューニングに従って $\mathcal{D}_{\text{retain}}$ 上の負の対数尤度を最小化する．

$$\mathcal{L}_{\text{retain}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{retain}}}\left[-\log p_\theta(x)\right]. \tag{3}$$

そして，式 2 と式 3 を統合し，特定能力の保持とそれ以外の排他的な忘却を実現する．

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{forget}}(\theta) + \lambda\cdot\mathcal{L}_{\text{retain}}(\theta), \tag{4}$$

ここで，正則化パラメータ λ は，忘却と保持のバランスを制御する．

3 実験設計

本節では，提案手法の有効性を検証するため，保持対象の知識・能力の保持と有害知識の忘却という二つの観点から包括的な評価を行う．

3.1 学習設計

医療と数学の応用を想定し，保持対象としてこれらの 2 つの Instruction Tuning データセットを用いる．具体的には，医療に AlpaCare [16] の MedInstruct-52k，数学に MetaMath [17] の MetaMathQA を使用する．Instruction Tuning データセットを選択した理由は，保持ドメインにおける多様な質問形式への対応能力を獲得するためである．

忘却には，モデル自身が生成した文章を用いる．具体的には BOS トークンを入力に合計 40,000 件（バッチサイズ 4，ステップ数 10,000）の文章を生成して，それを忘却する．生成の際は再現性を確保しつつ多様性も促すため，温度 2.0 を設定する．広範な知識の忘却には多様な長さの文章を用いることが重要であるため，本研究では 32，64，128，256 トークンの 4 種類の長さの文章を均等に混合する．学習には Llama-3.2-1B-Instruct を使用した．

3.2 評価設計

医療ドメインでは，AlpaCare [16] に従い，LLM による評価とベンチマーク評価の両方を行う．LLM による評価では，二つの医療データセットで自由形式の指示応答能力を評価する．第一のデータセットである MedInstruct-test は，AlpaCare で構築された評価用データセットであり，216

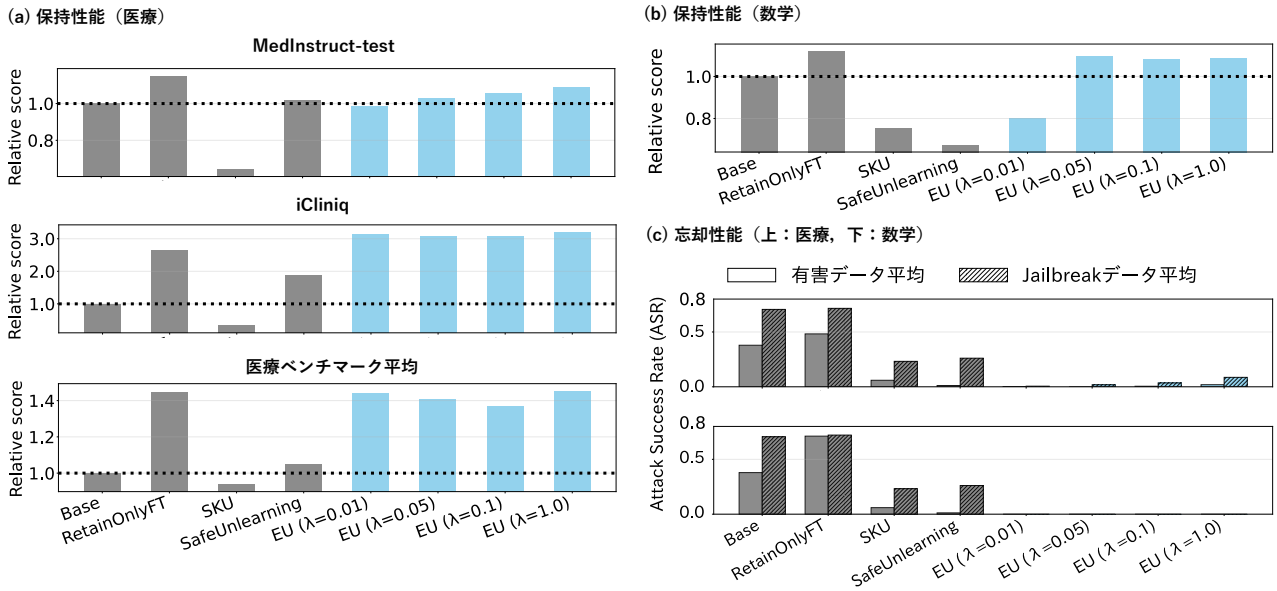

図 2　保持，忘却性能の定量評価．排他的逆学習（EU）を水色，その他ベースラインを灰色で表示している．

件の医療指示を含む．第二のデータセットである iCliniq は，ChatDoctor [18] で使用された実際の患者-医師会話の書き起こし 1000 件からなる．評価では，GPT-4o をジャッジモデルとして採用し，各テストケースについて評価対象モデルの応答と外部 LLM API による参照応答をペアワイズに比較する．包括的な評価のため，Text-davinci-003，GPT-3.5-Turbo，GPT-4，Claude-2 の 4 つの異なる API が生成した参照応答を用い，AlpacaEval [19] に従い勝率を計算し，その平均をとることでモデル性能をスコアリングする．ベンチマーク評価では，四つの医療多肢選択 QA（MedQA [20]，HeadQA [21]，PubmedQA [22]，MEDMCQA [23]）と要約データセットである MeQSUM [24] を用いて，多様な形式で評価する．lm-evaluation-harness [25] に従い，多肢選択 QA は正解率，要約タスクは ROUGE-L を計算する．

数学ドメインでは，MetaMath [17] と同様の GSM8K [26]，MATH [27] と，選択式 QA の MathQA [28] による評価を行う．lm-evaluation-harness に従い，生成タスクの GSM8K, MATH は EM（Exact Match）を，多肢選択式の MathQA は正解率を計算する．

忘却能力の評価では，有害データと Jailbreak 攻撃に対する防御性能を評価する．ベースの有害質問の評価には，GPTFuzzer [9] 由来の 100 件の有害質問と WildAttack [8] 由来の 217 件の有害質問を用いる．Jailbreak に対する防御性能の評価として，ベースの有害質問データセットに既存研究 [6] で公開されている 20 種類の Jailbreak プロンプトを組み合わせた，それぞれ 2000，4340 件のデータセットを用いる．これらは当然モデルの学習過程で未見である．Jailbreak プロンプトの詳細は § A.1 で示す．モデル出力の安全性を測るため，既存研究 [6] に倣い，ShieldLM-14B-qwen [29] をジャッジに用いて各データごとに Attack Success Rate (ASR) を計算する．ASR は安全でないと判断されたモデル応答の割合として定義される．

4　結果

保持，忘却性能の評価結果を図 2 に示す．ベースラインとして，元のモデル，保持対象データで通常のファインチューニングのみを行ったモデル（RetainOnlyFT）を用意し，これらのモデルと保持，忘却性能を比較する．また，既存の逆学習手法で未見の有害攻撃にも高い防御性能を示すことを報告している SKU [14]，SafeUnlearning [6] とも比較を行う．保持性能の評価では元のモデルのスコアに対する対象設定のスコアの比率を図 2 に可視化している．医療，数学ベンチマークはそれぞれの評価データでその比率の平均を計算している．また，排他的逆学習では，式 4 の λ でスケールし，0.01，0.05，0.1，1.0 の時の結果を記録している．

実験の結果，排他的逆学習（Exclusive Unlearning; EU）は医療，数学両方でベースモデルと同等以上の性能を保持し，データによっては RetainOnlyFT とも同等以上の性能を記録した．保持評価データには

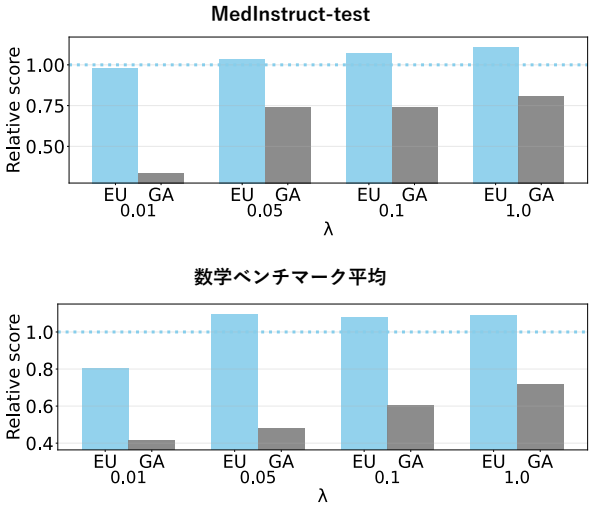

図 3　忘却項に一様損失を用いる排他的逆学習（EU）の設定と勾配上昇法による損失を用いる既存の設定（GA）の保持性能の比較.

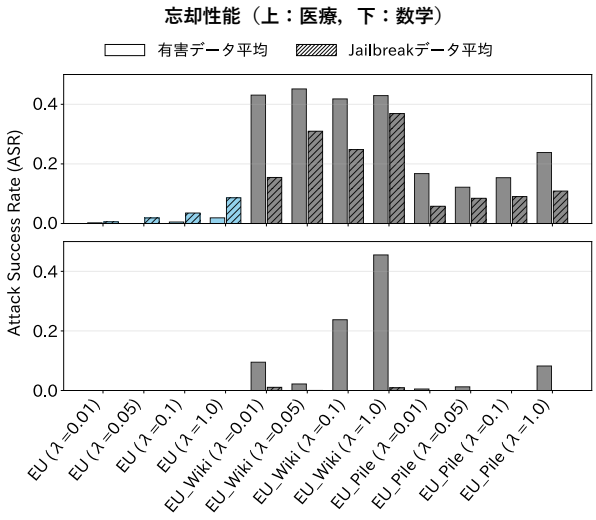

図 4　忘却にモデル自身が生成した文章を用いる場合と，大規模コーパス（Wiki, Pile）からサンプルする場合の攻撃成功率の比較.

§ 3.2 に示した通り，選択式タスク生成タスク両方の多様な形式を含んでおり，この結果は排他的逆学習後のモデルが保持領域内の広範な指示に答える能力を保持できていることを示す.

また，忘却性能の評価では，有害データや Jailbreak に対して，排他的逆学習後のモデルが既存の逆学習手法（SKU や SafeUnlearning）よりも強力な防御性能を持つことが示された.

5　議論

事前学習済みモデルから学習を行う妥当性　保持対象では有用な能力を保持した上で，対象外を排他的に忘却する目標を達成するために，事前学習済みモデルから出発することの必然性を示すため，保持データでランダム初期化モデルを事前学習する場合を検証した．その結果，ランダムパラメータからの出発では，§ A.2 のように言語生成能力が獲得できないため，評価結果はほとんど 0%になった．すなわち，本研究の目標を達成するためには，事前学習済みモデルから出発し，保持対象能力と言語生成能力を保持することが重要であることが示された.

忘却に一様損失を用いる妥当性　§ 2 で述べた通り，既存研究 [15] では，一様損失により忘却時に損失が有限値に収束することになるため，忘却学習が安定することが指摘されており，本研究ではより広範囲を忘却することになるため，この手法が必須であると考えた．この妥当性を示すため，忘却項に既存の Gradient Ascent（GA）を用いた場合と比較を行った．その結果を図 3 に示す．医療，数学両設定において，GA では学習前のモデルと比べ大きく保持性能が落ちることが確認でき，排他的逆学習に一様損失を使う妥当性が示された.

忘却にモデルの自己生成文章を使う妥当性　保持対象外の知識・能力を排他的に忘却するためにモデル自身が生成した文章を用いることの妥当性を示すため，Dolma コーパス [30] の Wikipedia サブセットからサンプルする設定（Wiki）や大規模コーパスである Pile [31] 全体からサンプルする設定（Pile）を比較ベースラインに設定し，学習設定を統一して比較した．図 4 から，忘却性能はモデル自身が生成した文章を用いる設定が最も高いことがわかる．さらに，巨大なコーパスからのランダムサンプリングは実務上困難であるという点を考慮しても，忘却にモデルの自己生成文章を使う妥当性が示される.

6　結論

既存の逆学習では，忘却すべき有害入力を網羅したデータを用意すること自体が難しく，忘却データに含まれない未見の有害質問への汎化が不足しがちである．本研究では，この課題に対し，保持すべき能力だけを残して他の能力を排他的に忘却する排他的逆学習を提案した．実験の結果，保持領域の性能を維持しながら，有害質問および Jailbreak に対する防御性能を示し，特に未見の Jailbreak に対する防御性能で既存手法を大きく上回ることを確認した.

謝辞

本研究成果は，データ活用社会創成プラットフォーム mdx を利用して得られたものであり，JST BOOST JPMJBY24A6，JPMJBY24B2 の支援を受けたものです．

参考文献

[1] S. Liu, et al. Rethinking machine unlearning for large language models. **Nature Machine Intelligence**, Vol. 7, No. 2, pp. 181–194, 2025.

[2] Jin Yao, et al. Machine unlearning of pre-trained large language models. In **ACL**, pp. 8403–8419, August 2024.

[3] Yao, et al. Large language model unlearning. In **NeurIPS**, Vol. 37, pp. 105425–105475, 2024.

[4] Yu Wang, et al. Large scale knowledge washing. In **ICLR**, Vol. 2025, pp. 74346–74366, 2025.

[5] Yaxuan Wang, et al. Llm unlearning via loss adjustment with only forget data. In **ICLR**, Vol. 2025, pp. 43076–43104, 2025.

[6] Zhexin Zhang, et al. From theft to bomb-making: The ripple effect of unlearning in defending against jailbreak attacks. **arXiv preprint:2407.02855**, 2025.

[7] Yi Liu, et al. Jailbreaking chatgpt via prompt engineering: An empirical study. **arXiv preprint:2305.13860**, 2024.

[8] Xinyue Shen, et al. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. **arXiv preprint:2308.03825**, 2024.

[9] Jiahao Yu, et al. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. **arXiv preprint:2309.10253**, 2023.

[10] Alexander Wei, et al. Jailbroken: How does llm safety training fail? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **NeurIPS**, Vol. 36, pp. 80079–80110, 2023.

[11] Andy Zou, et al. Universal and transferable adversarial attacks on aligned language models. **arXiv preprint:2307.15043**, 2023.

[12] Daniel Kang, et al. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. In **IEEE**, pp. 132–143, 2024.

[13] Xiaogeng Liu, et al. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In **ICLR**, 2024.

[14] Zheyuan Liu, et al. Towards safer large language models through machine unlearning. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 1817–1829, August 2024.

[15] Xiaojian Yuan, et al. A closer look at machine unlearning for large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, **ICLR**, Vol. 2025, pp. 49483–49508, 2025.

[16] Xinlu Zhang, et al. Alpacare:instruction-tuned large language models for medical application. **arXiv preprint:2310.14558**, 2023.

[17] Longhui Yu, et al. Metamath: Bootstrap your own mathematical questions for large language models. In **ICLR**, Vol. 2024, pp. 45040–45061, 2024.

[18] Yunxiang Li, et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. **arXiv preprint:2303.14070**, 2023.

[19] Xuechen Li, et al. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, May 2023.

[20] Di Jin, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. **arXiv preprint:2009.13081**, 2020.

[21] David Vilares, et al. HEAD-QA: A healthcare dataset for complex reasoning. In **ACL**, pp. 960–966, July 2019.

[22] Qiao Jin, et al. PubMedQA: A dataset for biomedical research question answering. In **EMNLP-IJCNLP**, pp. 2567–2577, November 2019.

[23] Ankit Pal, et al. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In **PMLR**, Vol. 174, pp. 248–260, 07–08 Apr 2022.

[24] Asma Ben Abacha and Dina Demner-Fushman. On the summarization of consumer health questions. In **ACL**, pp. 2228–2234, July 2019.

[25] Leo Gao, et al. The language model evaluation harness. https://github.com/EleutherAI/lm-evaluation-harness, July 2024.

[26] Karl Cobbe, et al. Training verifiers to solve math word problems. **arXiv preprint:2110.14168**, 2021.

[27] Dan Hendrycks, et al. Measuring mathematical problem solving with the math dataset. **NeurIPS**, 2021.

[28] Aida Amini, et al. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In **NAACL**, pp. 2357–2367, June 2019.

[29] Zhexin Zhang, et al. ShieldLM: Empowering LLMs as aligned, customizable and explainable safety detectors. In **EMNLP Findings**, pp. 10420–10438, November 2024.

[30] Luca Soldaini, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, et al., editors, **ACL**, pp. 15725–15788, August 2024.

[31] Leo Gao, et al. The Pile: An 800gb dataset of diverse text for language modeling. **arXiv preprint:2101.00027**, 2020.

表 1　本研究で扱った計 20 種類のジェイルブレイク攻撃の一覧．

攻撃タイプ	#	説明
ロールプレイ攻撃	4	モデルに善役と悪役を演じさせ，有害な内容を生成させる [7].
権限昇格攻撃	2	モデルに開発者モード等の制限解除モードを有効化させ，有害な内容を生成させる [7].
注意シフト攻撃	3	有害な問い合わせを一見無害な形式で包むことで，有害な応答を引き出す [7].
自動生成攻撃	8	手作業で作成した Jailbreak プロンプトを基に，Jailbreak プロンプトを自動生成する [9].
勾配攻撃	1	モデルの勾配を用いて敵対的プロンプトを最適化し，その最適化プロンプトを付加することで有害な応答を引き出す．Greedy Coordinate Gradient (GCG) [11] と呼ばれる手法である．
再整形攻撃	2	元の問い合わせの構造を変更する攻撃．例えばクエリを（a, b, c）のように分割し，a+b+c に答えるよう要求することで，有害な出力を誘発しうる [12, 13].

表 2　保持データでランダム初期化モデルを事前学習した際の医療，数学タスク応答例

保持対象	応答例
医療	the the the the the the the the the the the the the the the the the „„„„„„„„„„„„„„„„„„„„„„„„ ,,.,.,.,.,.,.,.,.,., and, and, and, and, and, and, and
数学	2.\nThe = 2 2.\nThe answer 2 = 2.\nThe 2.\nThe 2.\nThe 2.\nThe answer 2 2 = 2 = 2 = 3 2 = 3 3 3 3 3 3 3 3 2 = 3 3 3 3 3 3 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5

A　付録

A.1　Jailbreak プロンプトの詳細

実験に用いた 20 種類の Jailbreak プロンプトの詳細を表 1 に示す．既存研究 [6] と同様のプロンプトを用いており，6 タイプの多様な Jailbreak 手法を含んでいる．

A.2　事前学習済みモデルから学習を行う妥当性

表 2 に医療，数学を保持する設定において，保持データでランダム初期化モデルを事前学習する設定で実際に文章を生成した際の生成例を示す．医療データは MedInstruct の応答例，数学データは GSM8K の応答例である．いずれの場合も言語生成能力が構築されていない応答になっていることが確認できる．なお，このように意味不明な応答が続くため，表 2 では途中までの出力を示している．これは，保持対象のタスクを解く能力だけを残すという目的において，ランダム初期化モデルを保持データで事前学習するだけでは不十分であることを意味する．