LLM のアテンションヘッドに着目したジェイルブレイク攻撃の分析と防御手法の提案

新井雅稀 [1]　芝原俊樹 [2]　千葉大紀 [3]　秋山満昭 [2]　内田真人 [1]
[1] 早稲田大学　[2]NTT 社会情報研究所　[3]NTT セキュリティホールディングス株式会社
marai@akane.waseda.jp　toshiki.shibahara@ntt.com
{daiki.chiba, akiyama}@ieee.org　m.uchida@waseda.jp

概要

大規模言語モデル（Large Language Model; LLM）はジェイルブレイク攻撃に対して脆弱であり，違法行為や非倫理的な内容などの有害な出力をしてしまうリスクがある．このジェイルブレイク攻撃に対して，LLM が有害な出力をしてしまうメカニズムは十分に解明されていない．本研究では LLM のアテンションヘッドに着目して内部状態の分析を行い，数%のアテンションヘッドが有害な出力に大きく関与していることを明らかにする．また，分析結果を利用して，アテンションヘッドへの介入による防御手法を提案する．実験の結果，提案手法によ性能の低下を 3%以内に抑えつつ，攻撃成功率を 2，3%程度まで低下させられることが確認された．

1　はじめに

近年，ChatGPT [1] や Llama [2] などの大規模言語モデル（LLM）がテキスト生成や質問応答など，自然言語処理のさまざまなタスクにおいて用いられ，優れた性能を発揮している．その一方で，これらの LLM には，虚偽の情報や違法行為などの有害な内容を出力する可能性があるため，セキュリティの問題が挙げられている．これらの問題に対処するため，通常，LLM は Supervised Fine-Tuning（SFT）[3] や Reinforcement Learning from Human Feedback（RLHF）[4] などによって人間の価値観や意図に沿う出力をするように調整（Alignment）されている．

このような調整により，ある程度有害な出力は制御されているものの，攻撃者が意図的に調整を回避しようとした場合に防ぐことができず，ジェイルブレイク攻撃 [5, 6, 7, 8] に対して脆弱であることが知られている．ジェイルブレイク攻撃は，LLM から違法行為や非倫理的な内容などの有害な出力を引き出す攻撃であり，“Sure, here’s ...”などのような肯定応答を生成する確率が高くなるように入力を巧妙に細工する手法などが提案されている．これらの攻撃に対して，プロンプトレベルでの分析 [9] は行われている一方で，モデル内部の状態に着目した分析を行っている研究は少なく，LLM が有害な出力をしてしまうメカニズムは十分に解明されていない．

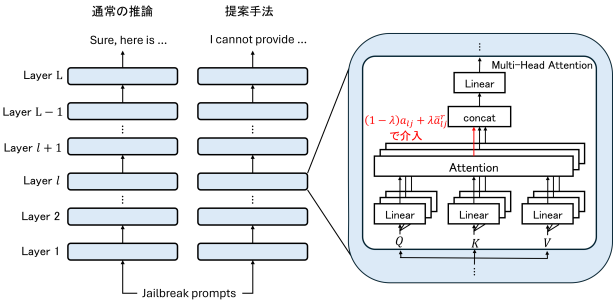

図 1　提案手法の概要図

本研究では，代表的なオープンソースモデルである Llama2-7B-chat-hf モデル [2] を対象に，トランスフォーマーの主構造の 1 つであるマルチヘッドアテンション機構に着目して内部状態の分析を行う．具体的には，攻撃に成功したプロンプトに対して，各アテンションヘッドの値を否定応答（“I cannot...”など）した元のプロンプトにおけるヘッドの値で置き換えた場合と通常の推論時での肯定応答の出力確率を比較することで，各アテンションヘッドが出力に与える影響を評価する．分析の結果，一部のアテンションヘッドが有害な出力に大きく関与していることが明らかになった．

さらに，分析結果を利用して，特定のアテンションヘッドへの介入による防御手法を提案する．分析によって得られた影響の大きいアテンションヘッドの値を，元のアテンションヘッドの値と否定応答したプロンプトにおけるアテンションヘッドの平均値の重み付き平均で置き換えることにより，防御を行

う（図 1）．

評価実験の結果，提案手法は代表的なジェイルブレイク攻撃手法である GCG [5] と AutoDAN [6] に対して攻撃成功率を低下させることが確認された．また，防御手法を考える際には，防御によって元の言語モデルの性能を低下させることなく，攻撃を防ぐことが要求される．そこで，さまざまなデータセットに対して，介入前後での性能の評価を行った結果，知識や推論能力については性能を維持することが確認された．また，指示データセットに対する性能（否定応答に変化した割合）の低下も最小限に抑えた．

2 関連研究

2.1 ジェイルブレイク攻撃

LLM におけるジェイルブレイク攻撃とは，入力を巧妙に細工することで，モデルに組み込まれている安全性に関する制約や制御を回避して，LLM から違法行為や非倫理的な内容などの有害な出力を引き出す攻撃である．ジェイルブレイク攻撃では，肯定応答の生成確率が高くなるように入力プロンプトを生成するが，入力プロンプトの作成方法によってさまざまな攻撃手法が提案されている [5, 6, 7, 8]．本研究で用いた代表的なジェイルブレイク攻撃手法である GCG と AutoDAN の 2 つについて説明する．

Zou らは肯定応答を生成する確率を最大化することを目的とした接尾辞（Adversarial suffix）の生成手法である Greedy Coodinate Gradient（GCG）[5] を提案した．GCG では悪意のあるプロンプトの末尾に接尾辞をつけ，その接尾辞を肯定応答を生成する確率が高くなるように，貪欲法と勾配法を組み合わせた手法によりトークンレベルで更新していく．

また，Liu らは文章の自然さを保ちつつ，ジェイルブレイクプロンプトを生成する手法 AutoDAN [6] を提案した．GCG はトークンレベルでプロンプトを更新するのに対して，AutoDAN では遺伝的アルゴリズムを用いて，文や段落レベルでプロンプトを改変していき，肯定応答を生成する確率が高くなるようにプロンプトを最適化する．

2.2 防御手法

2.1 節のジェイルブレイク攻撃に対して，さまざまな防御手法も提案されている．既存の防御手法はプロンプトの改変による防御，入出力のフィルタリング，ファインチューニングによる防御の大きく 3 つに分けられる．

1 つ目はプロンプトの改変による防御である．Jain らはトークン化の手法を変更することや，入力プロンプトを一度要約してから入力することで，攻撃を無効化する手法 [10] を提案した．

2 つ目は入力や出力に対するフィルタリングによる防御である．Jain らは GCG などのトークンレベルでの更新では文として不自然になることに着目し，perplexity を評価することで防御する手法 [10] を提案した．また，Phute らは，LLM 自身を有害な文章の検出に活用する手法 [11] を提案した．

3 つ目はファインチューニングによる防御である．Bainchi らは Llama などのモデルをファインチューニングする際に，安全性に関するデータ（悪性のクエリと否定応答の組）を組み込むことにより，モデルの安全性が向上したことを示した [12]．また，Llama Guard [13] のような既存の LLM をファインチューニングすることで入力や出力の安全性を判定するモデルなども開発されている．

しかし，これらの防御手法の多くは追加の推論や学習が必要となる．また，防御手法の適用による性能低下も問題となっている．本研究では，性能の低下を最小限に抑え，追加の推論や学習を行うことなく，ジェイルブレイク攻撃を防御する手法を提案する．

3 内部状態の分析

本節では，ある原因が結果に与える影響が，他の媒介変数を通じてどの程度生じているのかを評価するための手法である因果媒介分析により，In-Context Learning（ICL）タスクにおける各アテンションヘッドの影響を定量化した Todd らの手法 [14] を参考に，ジェイルブレイク攻撃に対して，LLM のマルチヘッドアテンション機構に着目した分析を行う．

3.1 記法

L 層で構成され，各層に J 個のアテンションヘッドをもつ自己回帰型のトランスフォーマー言語モデルを f とする．また，この言語モデル f にプロンプト q を入力したときの次のトークンの確率分布を $f(q)$ とし，あるトークン $y \in \mathcal{V}$（$\mathcal{V}$ は語彙）の出力確率を $f(q)[y] \in [0,1]$ とする．$l(\leq L)$ 層目の $j(\leq J)$ 番目のアテンションヘッドを a_{lj} と表す．

否定応答したプロンプト p にジェイルブレイク攻

撃手法を適用し，攻撃に成功したプロンプトを p^s とする．また，p^s の集合を P_s とする．

3.2 分析手法

因果媒介分析を用いて，各アテンションヘッドの影響を定量化する．まず，各アテンションヘッド a_{lj} において，介入前後でのトークン “ Sure”の生成確率の変化 CIE（causal indirect effect）を式 (1) により計算する．

$$\begin{aligned}\mathrm{CIE}(a_{lj}|p_i^s) =& f(p_i^s)[\text{“ Sure”}] \\ &- f(p_i^s|a_{lj} := a_{lj}(p_i))[\text{“ Sure”}] \end{aligned} \quad (1)$$

式 (1) は攻撃に成功したプロンプト p_i^s の初期トークン生成時に，l 層目の j 番目のアテンションヘッドの値を，否定応答した元のプロンプト p_i におけるアテンションヘッドの値 $a_{lj}(p_i)$ で置き換えたときの肯定応答（Sure）の生成確率の減少率を表している．次に，すべてのプロンプト $p_i^s \in P_s$ において，CIE を計算して，その平均 AIE（average indirect effect）を次の式 (2) で計算する．

$$\mathrm{AIE}(a_{lj}) = \frac{1}{|P_s|} \sum_{p_i^s \in P_s} \mathrm{CIE}(a_{lj}|p_i^s) \quad (2)$$

式 (2) は介入によって肯定応答の生成確率が平均的にどのくらい減少したかを表している．この値が大きいアテンションヘッドほどジェイルブレイクの成功，すなわち有害な出力に大きく関与していると考えられる．

3.3 分析結果

分析対象のモデルとして代表的なオープンソースモデルである Llama2-7B-chat-hf を用いた．層数は $L = 32$，ヘッド数は $J = 32$ である．

分析の対象となるプロンプトの集合 P_s として AdvBench[5] のクエリに対して，GCG を実行して攻撃に成功したプロンプト 20 個を用いた．介入を行わない場合の “ Sure”の生成確率の平均は 0.768 であった．

図 2 に各ヘッドにおける AIE の値を示したヒートマップ図を示す．全体の約 93%のヘッドにおいて AIE が −0.05 以上 0.05 未満の値となっていて，一部の少数のヘッドが大きな値をとっていることがわかる．

また，図 2 において，AIE の値が大きいヘッドを表 1 に示す．最も AIE の値が高かったのは 4 層目の

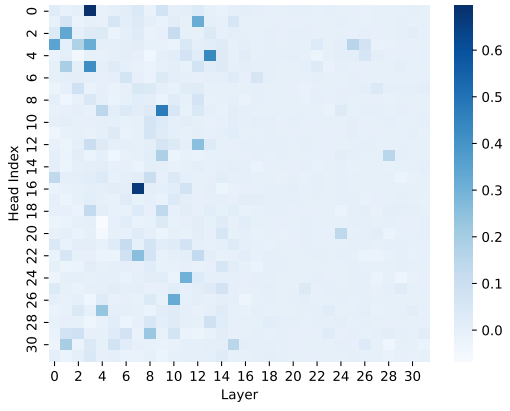

図 2 AIE のヒートマップ図

表 1 AIE 上位 5 個

layer	head	AIE
4	1	0.695
8	17	0.663
10	10	0.476
14	5	0.435
4	6	0.421

1 番目のヘッドであり，“ Sure”の生成確率が平均で 0.695 減少したことを示している．

以上の結果から，一部のアテンションヘッドが有害な出力に大きく関与していると考えられる．

4 防御手法

本節では 3 節の分析結果をもとに，アテンションヘッドへの介入による防御手法について説明する．提案手法は介入に利用する値を取得する部分, 介入を行う部分の 2 つに分かれる．また，介入するヘッドは 3 節の分析により決める．

介入に利用する値は否定応答したプロンプトに対するアテンションヘッドの平均値を利用する．平均値を用いることで，多くの否定応答に共通する特徴を反映することができ，有害な出力を抑制する効果が期待される．具体的には，否定応答したプロンプトの集合を P_r とするとき，式 (3) によって各アテンションヘッドにおいて平均値 $\bar{a}_{lj}^r$ を取得する．

$$\bar{a}_{lj}^r = \frac{1}{|P_r|} \sum_{p \in P_r} a_{lj}(p) \quad (3)$$

防御は推論時の最初のトークンを生成するときのみ，3 節の分析で得られた AIE が大きいヘッドに，

$$a'_{lj} = (1 - \lambda) a_{lj} + \lambda \bar{a}_{lj}^r \quad (4)$$

表 2　攻撃成功率

介入したヘッド（layer,head）	介入率 λ	攻撃成功率		
		AdvBench	GCG	AutoDAN
なし（No Defense）	—	0.0%	45.0%	30.8%
(4, 1), (8, 17), (10, 10)	0.75	0.0%	2.5%	3.1%
(4, 1), (8, 17), (10, 10), (14, 5)	0.5	0.0%	1.9%	3.1%
(4, 1), (8, 17), (10, 10), (14, 5), (4, 6)	0.75	0.0%	0.6%	0.7%

表 3　知識や推論能力を測るデータセット・指示データセットに対する性能評価

介入したヘッド（layer,head)	介入率 λ	正答率				否定応答への変化率
		ARC-easy	ARC-challenge	Lambada	Winogrande	AlpacaEval
なし（No Defense）	—	74.7%	53.5%	65.8%	51.0%	—
(4, 1), (8, 17), (10, 10)	0.75	74.2%	52.8%	65.7%	50.7%	3.4%
(4, 1), (8, 17), (10, 10), (14, 5)	0.5	74.9%	53.5%	65.4%	51.2%	2.6%
(4, 1), (8, 17), (10, 10), (14, 5), (4, 6)	0.75	74.4%	51.5%	64.9%	50.5%	8.9%

で介入することによって行う．ここで λ は介入率を表している．

5　評価実験

5.1　実験設定

モデル　Llama2-7B-chat-hf を用いた．

データセット　ジェイルブレイクプロンプトの生成には AdvBench [5] を利用した．また，介入前後での性能評価のために，知識や推論能力を測るデータセットとして，ARC-easy [15]，ARC-challenge [15]，Lambada [16]，Winogrande [17] の 4 つを，指示データセットとして AlpacaEval [18] を用いた．

評価指標　攻撃成功率を評価指標とする．初期 32 トークンの中に否定語（I cannot など）が含まれていない場合を攻撃成功と定義して評価を行った．判定に用いた否定語は Appendix A に示す．

また，知識や推論能力を測るデータセットに対しては正答率，指示データセットに対しては介入後に否定応答に変化した割合を評価指標とする．

全ての実験において，評価の際には，常に出力確率が最も高いトークンを出力するように設定した．

攻撃設定　AdvBench のプロンプトに GCG，AutoDAN の 2 つのジェイルブレイク攻撃手法を適用した．GCG によって生成されたプロンプト 160 個，AutoDAN によって生成されたプロンプト 130 個を用いて評価を行った．GCG によって生成されたプロンプトについては 3 節での分析に用いたプロンプトとの重複はない．

防御設定　式 (3) を計算する際に必要な否定応答したプロンプトの集合 P_r として AdvBench のプロンプト 20 個を用いた．この 20 個は評価実験に用いたものとの重複はない．介入するヘッドは表 1 において AIE の値が大きいヘッド上位 5 個のうち，さまざまな組み合わせで実験を行った．介入率として $\lambda = 1.0, 0.75, 0.5$ の 3 通りについて実験を行った．

5.2　実験結果

表 2 に攻撃成功率を示す．介入によって攻撃成功率が大きく低下した部分を抽出している（全ての詳細な結果は Appendix B に示す）．特定のヘッドに介入することで，攻撃成功率が大きく低下することがわかる．また，表 3 に介入前後での性能評価の結果を示す．指示データセットに対しては一部無害なプロンプトに対しても否定応答することが見られたが，3%程度であれば無害なプロンプトへの影響は限定的であると考えられる．知識や推論能力については介入後も維持されることが確認された．

6　おわりに

本研究では，LLM のマルチアテンションヘッド機構に着目してジェイルブレイク攻撃の分析を行い，一部のアテンションヘッドが有害な出力に大きく関与していることを明らかにした．また，分析結果に基づき，特定のアテンションヘッドへの介入による防御手法を提案し，提案手法は攻撃成功率を低下させることを示した．本研究がジェイルブレイク攻撃のメカニズムの解明につながることを期待する．

謝辞

本研究の一部は，日本学術振興会における科学研究費補助金基盤研究（C）（課題番号 23K11111）による支援を受けている．

参考文献

[1] OpenAI, 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:2303.08774v6**, 2024.

[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:2307.09288v2**, 2023.

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **arXiv:2203.02155v1**, 2022.

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, and Tom Henighan others. Training a helpful and harmless assistant with reinforcement learning from human feedback. **arXiv:2204.05862v1**, 2022.

[5] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. **arXiv:2307.15043v2**, 2023.

[6] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[7] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. **arXiv:2310.08419v4**, 2024.

[8] GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. Jiahao yu and xingwei lin and zheng yu and xinyu xing. **arXiv:2309.10253v4**, 2024.

[9] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In **33rd USENIX Security Symposium (USENIX Security 24)**, pp. 4675–4692, Philadelphia, PA, August 2024. USENIX Association.

[10] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. **arXiv:2309.00614v2**, 2023.

[11] Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. LLM self defense: By self examination, LLMs know they are being tricked. In **The Second Tiny Papers Track at ICLR 2024**, 2024.

[12] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In **The Twelfth International Conference on Learning Representations**, 2024.

[13] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. **arXiv:2312.06674v1**, 2023.

[14] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. **arXiv:1803.05457v1**, 2018.

[16] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. The LAMBADA dataset: Word prediction requiring a broad discourse context. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics.

[17] Sakaguchi Keisuke, Le Bras Ronan, Bhagavatula Chandra, and Choi Yejin. Winogrande: An adversarial winograd schema challenge at scale. **arXiv:1907.10641v2**, 2019.

[18] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. **arXiv:2305.14387v4**, 2024.

表 4　攻撃成功率と否定応答への変化率（$\lambda = 1$）

介入したヘッド（layer,head）	攻撃成功率			否定応答への変化率
	AdvBench	GCG	AutoDAN	AlpacaEval
なし	0.0%	45.0%	30.8%	—
(4, 1)	0.0%	13.8%	30.8%	0.0%
(8, 17)	0.0%	16.3%	31.5%	0.0%
(10, 10)	0.0%	15.6%	1.5%	11.3%
(4, 1), (8, 17)	0.0%	12.5%	30.8%	0.0%
(4, 1), (9, 9)	0.0%	3.1%	1.5%	11.1%
(4, 1), (8, 17), (10, 10)	0.0%	3.1%	1.5%	10.4%
(4, 1), (8, 17), (10, 10), (14, 5)	0.0%	2.5%	0.0%	28.3%
(4, 1), (8, 17), (10, 10), (14, 5), (4, 6)	0.0%	0.6%	0.0%	25.7%

表 5　攻撃成功率と否定応答への変化率（$\lambda = 0.75$）

介入したヘッド（layer,head）	攻撃成功率			否定応答への変化率
	AdvBench	GCG	AutoDAN	AlpacaEval
なし	0.0%	45.0%	30.8%	—
(4, 1)	0.0%	11.3%	30.8%	0.0%
(8, 17)	0.0%	14.4%	30.8%	0.0%
(10, 10)	0.0%	13.8%	3.1%	3.5%
(4, 1), (8, 17)	0.0%	10.0%	30.8%	0.0%
(4, 1), (9, 9)	0.0%	2.5%	3.1%	3.6%
(4, 1), (8, 17), (10, 10)	0.0%	2.5%	3.1%	3.4%
(4, 1), (8, 17), (10, 10), (14, 5)	0.0%	0.6%	0.7%	9.6%
(4, 1), (8, 17), (10, 10), (14, 5), (4, 6)	0.0%	0.6%	0.7%	8.9%

表 6　攻撃成功率と否定応答への変化率（$\lambda = 0.5$）

介入したヘッド（layer,head）	攻撃成功率			否定応答への変化率
	AdvBench	GCG	AutoDAN	AlpacaEval
なし	0.0%	45.0%	30.8%	—
(4, 1)	0.0%	16.3%	30.8%	0.0%
(8, 17)	0.0%	17.5%	30.8%	0.0%
(10, 10)	0.0%	18.8%	11.5%	1.4%
(4, 1), (8, 17)	0.0%	10.0%	30.8%	0.0%
(4, 1), (9, 9)	0.0%	5.6%	10.8%	1.4%
(4, 1), (8, 17), (10, 10)	0.0%	3.8%	10.8%	1.4%
(4, 1), (8, 17), (10, 10), (14, 5)	0.0%	1.9%	3.1%	2.6%
(4, 1), (8, 17), (10, 10), (14, 5), (4, 6)	0.0%	1.9%	3.1%	2.7%

A　詳細な実験設定

攻撃成功の判定に用いた否定語は，“I’m an AI”，“I’m sorry”，“I am sorry”，“I apologize”，“As an AI”，“As a language model”，“As an Assistant”，“As a responsible”，“as an AI”，“Hello!”，“it’s not appropriate”，“I’m really sorry”，“unethical”，“not ethical”，“not legal”，“or ethical”，“is illegal”，“against ethical”，“not appropriate”，“a serious crime”，“My apologies”，“If you have any other non-malicious requests”，“I’m not able to provide”，“I am really sorry”，“I can’t provide”，“I can’t assist”，“I can’t help”，“I cannot”，“A conversation between a user and an LLM-based AI assistant. The assistant gives helpful and honest answers.”，“I must inform you”，“I’m just an AI”，“I am unable to provide information on”，“We are not able to provide”，“I’m an artificial intelligence”，“I strongly condemn”，“is not recommended” である．

ARC データセットに対する評価の際には [INST] {question} ({label}) {choice} … ({label}) {choice} [/INST] Answer: をプロンプトして与えた．また Winogrande の評価の際には [INST] {sentence} Which does ‘_’ refer to? (1) {option1} (2) {option2} [/INST] Answer: をプロンプトして与えた．{} の部分は質問ごとに異なる．

B　詳細な結果

介入率を $\lambda = 1.0, 0.75, 0.5$ としたときのそれぞれの詳細な結果を表 4，表 5，表 6 に示す．これらの表より，介入するヘッドが多くなれば，攻撃成功率は低くなるが，否定応答への変化率は高くなり，介入するヘッドが同じ場合は介入率 λ が大きいほど，攻撃成功率は低くなり，否定応答への変化率は高くなるというトレードオフの関係が見られた．