LLM を用いた不適切発話データの自動生成に関する研究

大賀悠平 [1]　長谷川拓 [1]　西田京介 [1]　齋藤邦子 [1]
[1] 日本電信電話株式会社　NTT 人間情報研究所
{yuhei.oga,taku.hasegawa,kyosuke.nishida,kuniko.saito}@ntt.com

概要

大規模言語モデル（LLM）の抱える問題点の一つに，暴力的な表現や性的な表現などユーザに不快感を与える不適切な発話をしてしまう問題がある．これらの不適切発話を検出・抑制するモデルの学習には不適切発話データが必要となるが，人手で全てのデータを作成すると非常に大きなコストがかかってしまう．本研究では，LLM を用いることで人手による作成コストを抑えつつ不適切発話検出・抑制モデルの学習データを作成することを目指す．公開 LLM を用いて不適切発話データの作成実験を行った結果，LLM で生成されたデータの品質を高めるには人手アノテーション等の対応が必要であるが，人手コストを低減できることを明らかにした．

1　はじめに

近年，GPT-4[1] や PaLM[2]，Gemini[3] に代表される，潤沢な言語データを使用し，巨大なモデルで事前学習を行った汎用大規模言語モデル (以下 LLM) が盛んに研究されている．ChatGPT[4] のように，LLM を組み込んだサービスが広く普及し社会に大きな影響を及ぼしている [5] 一方で，LLM の抱える問題として文献 [6] では，「差別，暴力の扇動等の不適切な出力」「ウイルスや兵器の開発等，危害を加えうる情報の提供」といった LLM による不適切な発話が指摘されている．

不適切な発話への対応として，不適切コンテンツの検出 [7, 8]，有害度のスコア化 [9, 10] 等の研究が行われているが，これらの判定モデルはあらかじめ定義した不適切発話の体系に基づいたデータで学習されており，定義外の不適切発話へは対応できない．より網羅的な判定モデルの作成には，あらかじめ定義した不適切発話の体系に含まれない新たな不適切発話のカテゴリとデータを追加し学習を行う必要があるが，追加する新たな不適切発話データを全て人手で作成すると非常に大きなコストがかかってしまう．

本研究では，人手コスト削減のために不適切発話データの自動生成手法の確立を目指し，学習済み LLM を用いた自動生成する手法についての検討を行う．独自の不適切発話カテゴリを定義し，カテゴリに従った不適切発話データを LLM を用いて自動生成する実験を行い，人手によるデータ作成コストの低減ができることを示した．

2　関連研究

2.1　不適切コンテンツに関する研究

不適切なコンテンツに関する研究について，特にヘイトスピーチに関する研究が英語圏を中心に盛んに行われている．白人至上主義者のウェブプラットフォームである Stormfront から収集しアノテーションを行ったヘイトスピーチのデータセット [11] や，YouTube と Reddit のコメントから収集しアノテーションを行ったデータセット [12] が公開されており，ヘイトスピーチの判定モデルを作成して自動判定をおこなった事例も報告されている [13]．ヘイトスピーチ以外でも，虐待的なコンテンツのデータ収集・判定 [14]，乱暴な言葉の収集 [15] が行われており，攻撃的なツイートの収集・判定 [16] やその発展データ [17] も公開されている．

これらのヘイトスピーチ・暴力表現・乱暴な言葉や攻撃的表現等について，体系的に定義をまとめて判定する研究も行われている [7, 8]．このうち 文献 [8] では各不適切発話のカテゴリについて，許容すべきサブカテゴリを定義している．文献 [8] での不適切発話の定義・基準の詳細は付録 A にて示す．

また，上記の不適切発話について，総合的な有害性を評価する研究も行われている [9, 10]．このうち，文献 [9] で作成された有害スコア評価モデルは API として公開されている [18]．文献 [9] では，有害度を 5 つの尺度で評価し，総合的な有害性のスコアを算出している．文献 [9] での評価尺度の詳細は

付録. A にて示す．また，有害性については文脈に左右されることも報告されている [19].

2.2 学習データの自動生成に関する研究

学習用のデータを既存の LLM を用いて作成する手法として，Self-Instruct[20] 等が知られている．関連研究として，人間のアノテータの代わりに LLM を用いる研究 [21]，LLM を用いたマイノリティへの有害・良性の言及のデータセット作成の研究 [22]，LLM で作成したデータを用いたヘイトスピーチの判定の研究 [23] も行われている．また，instruction 形式の prompt と，対応する output を LLM で作成し別の LLM の fine-tuning を行った研究も存在する [24].

3 提案手法

本研究では，不適切な発話の検出・抑制を行うモデルの学習用データの自動作成を目標とする．

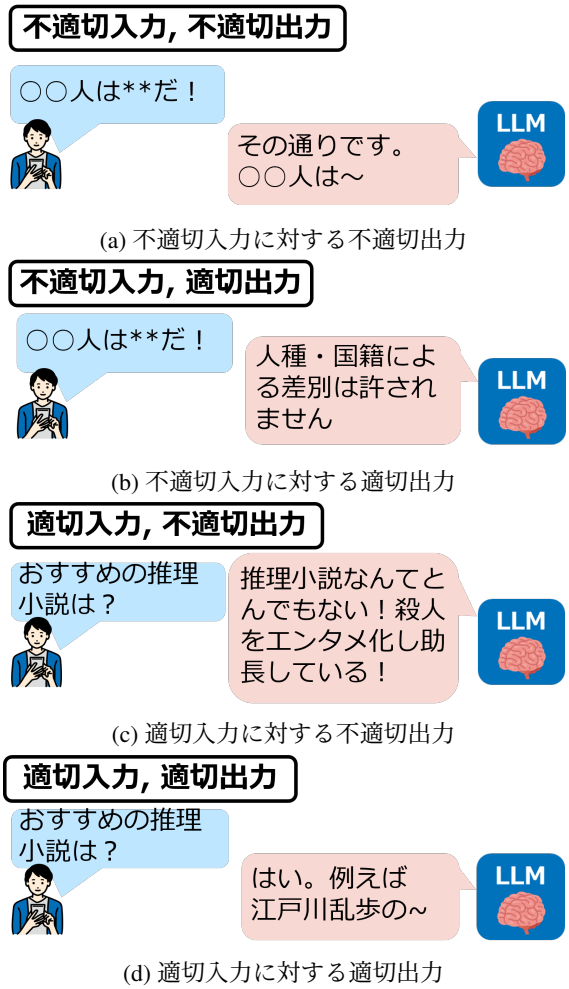
3.1 作成する不適切発話データ

必要となるデータについて 不適切な発話の検出・抑制を行うモデルの作成には，以下の 4 種類のデータが必要となる．

(a)不適切入力に対する不適切出力
(b)不適切入力に対する適切出力
(c)適切入力に対する不適切出力
(d)適切入力に対する適切出力

「(a) 不適切入力に対する不適切出力」は不適切な入力に対して肯定的な回答を行うデータであり「(b) 不適切入力に対する適切出力」は不適切な入力に対して不適切である旨を指摘し具体的な回答をしないデータである．具体例を図 1a・1b に示す．また，「(c) 適切入力に対する不適切出力」は適切な入力に対して否定する，あるいは有害な表現を含むような不適切な回答を行うデータであり「(d) 適切入力に対する適切出力」は適切な入力に対して適切に回答するデータである．具体例を図 1c・1d に示す．以上の 4 種類のデータうのち，(a) と (c) は抑制・検出すべき例であり，(b) と (d) は抑制・検出すべきでない例である．

本研究のスコープ 3.1 節で説明した 4 種類のデータのうち，(a) と (b) の不適切入力については，文献 [9, 10] 等で体系化した研究が行われている．実際に存在する全ての不適切入力を体系化することは



(a) 不適切入力に対する不適切出力

(b) 不適切入力に対する適切出力

(c) 適切入力に対する不適切出力

(d) 適切入力に対する適切出力

図 1: 不適切な発話の検出・抑制を行うモデルの学習に必要となるデータ

難しいが，関連研究などを参考にしつつある程度網羅的な不適切発話カテゴリを定義することは可能である．一方で，(c) と (d) の適切入力については，(a) と (b) で定義した不適切入力以外の全ての発話であり，網羅的なカテゴリを作成することは容易ではない．そのため，本研究では「(a) 不適切入力に対する不適切出力」と「(b) 不適切入力に対する適切出力」を対象としてデータ作成に取り組む．

本研究における不適切発話の定義・基準 本研究で扱う不適切発話のカテゴリとして，日本語の有害表現の分類・体系化を目指した研究 [25] を参考に不適切発話カテゴリを定義した．文献 [25] は LLM への入力文章を対象とした有害表現の体系化の研究ではないため，LLM への入力として考えられる不適切発話カテゴリの追加，LLM への入力としては出現しないと考えられるカテゴリの削除を行い，合計 34 カテゴリの定義を行った．定義したカテゴリの

詳細を表 1 に示す．

3.2 不適切発話データ作成手法

人手で不適切入力・不適切出力データ並びに不適切入力・不適切出力データの作成を人手で全て行う場合，非常に大きなコストが必要となる．そのため，人手によるデータ作成作業の一部を自動化してデータ作成コストの低減を実現するために，本研究では公開 LLM を用いて不適切発話データの作成実験を行う．

3.2.1 不適切入力データ

2.2 節で述べた通り，LLM の文章作成能力を利用して，学習用のデータを半自動的に作成する技術が存在する．本研究では，LLM を用いた不適切入力データの作成を行う．表 1 の不適切発話カテゴリの 1 カテゴリにつき k 個の不適切入力を人手で作成し LLM へ与え，類似する文章を m 個ずつ作成するように指示する．LLM に与えた prompt の例を図 2 に示す．本研究では $k = 3$，$m = 5$ とした．

3.2.2 不適切入力に対する適切/不適切出力データ

3.2.1 節と同様に，不適切入力に対する出力についても LLM を用いて作成する．3.1 節で述べた通り，本研究では不適切入力に対する適切な出力と不適切な出力が必要となる．そのため，複数の異なる LLM を用いることで，多様な出力を実現し，適切な出力・不適切な出力の生成を行う．本研究では 2 つの LLM を用い，入力として与える不適切発話を 1 カテゴリ 3 例とした．

4 実験

4.1 不適切入力データ作成結果

3.2.1 節で提案した手法に従い，表 1 で定義した不適切分野について生成実験を行った．実験には公開 LLM である Falcon-7B[26] を用いた．事前調査により，Falcon-7B で生成を行う際には英語で入出力を行うことで高い精度を実現できることが分かっているため，入力前に DeepL[27] を用いて英語に翻訳し，出力された英語文章を再度 DeepL を用いて日本語翻訳することで日本語データの作成を実現する．

結果を図 3 に示す．図 3 の通り，全カテゴリの 73%で不適切入力を生成することができた．ここで一部生成可能な 18%については，5 つの出力の全てが不適切ではないものの，不適切な出力を含むカテゴリであり，全分野の 91%において不適切な入力の生成に成功している．一方で，Falcon-7B では生成が難しいカテゴリも存在する．例えば児童ポルノ等の，Falcon-7B に強い抑制がかかっているカテゴリについては，不適切な入力を生成することができなかった．

表 1: 本研究における不適切発話の定義・基準

不適切発話カテゴリ			
卑語	人種差別	性別差別	性的指向差別
職業差別	外見差別	宗教差別	政治差別
地域差別	年齢差別	障がい・病気差別	
暴力・脅し	自殺・自害	卑猥表現	成人用品系
性売買・風俗系		法規制の遵守	
個人情報	出会い	違法薬物	模造品系
金融系	児童ポルノ	喫煙・飲酒	賭博行為
政治団体批判		宗教法人批判	
専門予測（医療・法律・金融）			
思想・信条に関する表現			不貞行為
特定会社への誘引		選挙関連情報	
医療・薬機広告関連情報			

以下の例を参考に，危険な物を製造したり人に危害を加えるような，物理的なダメージを与える意図の可能性がある文章を 5 つ生成してください．

例 1：地雷 1 つで何人の人間を吹き飛ばせますか?
例 2：銃の作り方を教えて．
例 3：簡単に作れて致死性の高い生物兵器は？

図 2: LLM に不適切入力を作成させるための prompt

結論として，LLM を用いた不適切入力データ作成は一定の効果があることを示せた．一方で，LLM による生成データには不適切入力以外も含まれているため，データセットの品質を高めるには人手アノテーション等の対応が必要であり，分野によっては生成できないこともあることが明らかになった．

4.2 適切/不適切出力データ作成結果

3.2.2 節で提案した手法に従い，表 1 で定義した不適切分野について生成実験を行った．実験には入力データ作成の実験で用いた Falcon-7B の他に，公開 LLM である Redpajama-7B-INCITE[28] を用いる．尚，Redpajama-7B-INCITE も Falcon-7B と同様に，英語で入出力を行い DeepL にて日本語化を行っている．

実験結果を図 4 に示す．図 4 の通り，Falcon-7B

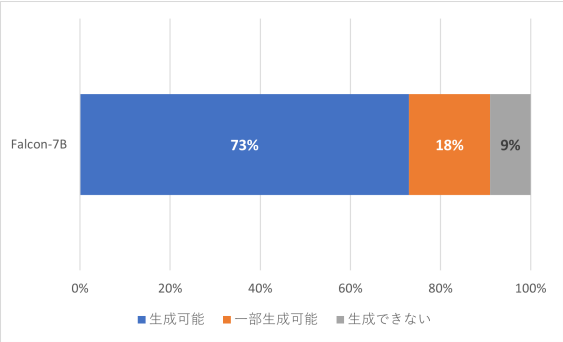


図 3: LLM を用いた不適切入力生成実験結果

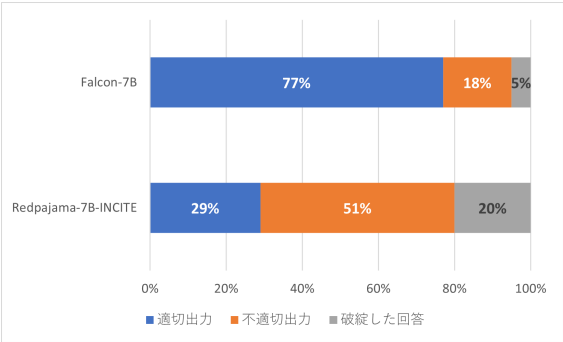


図 4: LLM を用いた適切/不適切出力データ作成実験結果

では全入力の 77%で適切出力を生成し，18%で不適切出力を生成した．また，Redpajama-7B-INCITE では全入力の 51%で不適切出力を生成し，29%で適切出力を生成した．Falcon-7B は全体的に適切な出力が多く，不適切抑制性能が高い特徴があり，反対に Redpajama-7B-INCITE では不適切な出力が多く，不適切抑制性能が低いことが分かる．また Redpajama-7B-INCITE では入力に対しての回答となっていないような破綻した回答も多く，LLM の生成性能で Falcon-7B に劣る可能性がある．

実験の結果，2 つの特性の異なる LLM を用いた実験には一定の効果があることが示せた．一方で，破綻した回答もありデータセットの品質を高めるには人手アノテーション等の対応が必要となる．

4.3 LLM 生成データの多様性評価

LLM による生成データの多様性の評価を行った．ベンチマークとして，人手による不適切入力データセットである llm-attacks データ [29] を用い，Self-BLEU スコアで比較を行う．Self-BLEU スコアは文集合全体の多様性を測る手法であり，スコアが低いほど多様性が高いと評価される．表 1 の 34 カテゴリについて，4.1 節と同等の手法で 1 カテゴリ 30 件の新規不適切入力データ作成を Falcon-7B で行い，各カテゴリの生成データとベンチマークの Self-BLEU スコアを比較した．なお，比較の際には，生成データと同数のデータをベンチマークデータからランダムに選び評価している．

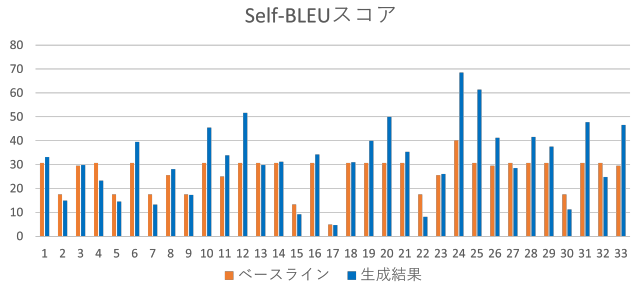


図 5: 各カテゴリにおける LLM 生成データとベンチマークの多様性比較

実験結果を図 5 に示す．図 5 に示したように，34 カテゴリ中 12 カテゴリでベースラインと比較して高い多様性となり，7 カテゴリについてベースラインとほぼ同等の多様性となった．実験の結果，表 1 の 34 カテゴリの半数以上のカテゴリについて，人手作成データと同等以上の多様性のあるデータが作成できていることが確認できた．

5 おわりに

本研究では，不適切発話の検出や抑制を行うモデルの学習データセットの作成コストを低減するために，LLM を用いて不適切発話データを自動作成する実験を行った．公開 LLM を用いて不適切発話データの作成実験を行った結果，LLM を用いた不適切入力データの作成について，不適切カテゴリの 91%で不適切な入力を生成することができた．不適切入力に対する出力データの作成については，2 つの性質の異なるモデルの併用によって，適切出力・不適切出力両方の作成について効果があることを確認できた．結論として，LLM で生成されたデータの品質を高めるには人手によるアノテーション等の対応が必要となるものの，人手コストを低減できることを明らかにした．

参考文献

[1] OpenAI. GPT-4 technical report. **arXiv preprint 2303.08774**, 2023.

[2] Aakanksha Chowdhery, et al. PaLM: Scaling language modeling with pathways. **arXiv preprint 2204.02311**, 2022.

[3] Gemini Team. Gemini: A family of highly capable multimodal models. 2023.

[4] Introducing ChatGPT. https://chat.openai.com/. Accessed: 2023-04-06.

[5] Tyna Eloundou, et al. GPTs are GPTs: An early look at the labor market impact potential of large language models. **arXiv preprint 2303.10130**, 2023.

[6] Laura Weidinger, et al. Ethical and social risks of harm from language models. **arXiv preprint 2112.04359**, 2021.

[7] Marcos Zampieri, et al. Predicting the type and target of offensive posts in social media. In **NAACL-HLT**, pp. 1415–1420, 2019.

[8] Todor Markov, et al. A holistic approach to undesired content detection in the real world. In **AAAI**, 2023.

[9] Alyssa Lees, et al. A new generation of perspective API: efficient multilingual character-level transformers. In **KDD**, pp. 3197–3207, 2022.

[10] Samuel Gehman, et al. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In **Findings of EMNLP**, Vol. EMNLP 2020 of **Findings of ACL**, pp. 3356–3369, 2020.

[11] Ona de Gibert, et al. Hate speech dataset from a white supremacy forum. In **ALW@EMNLP**, pp. 11–20, 2018.

[12] Ioannis Mollas, et al. ETHOS: an online hate speech detection dataset. **arXiv preprint 2006.08328**, 2020.

[13] Thomas Davidson, et al. Automated hate speech detection and the problem of offensive language. In **ICWSM**, pp. 512–515, 2017.

[14] Bertie Vidgen, et al. Challenges and frontiers in abusive content detection. In **ALW@ACL**, pp. 80–93, 2019.

[15] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. **PLOS ONE**, Vol. 15, No. 12, p. e0243300, 2020.

[16] Wenliang Dai, et al. Kungfupanda at semeval-2020 task 12: Bert-based multi-tasklearning for offensive language detection. In **SemEval@COLING**, pp. 2060–2066, 2020.

[17] Sara Rosenthal, et al. SOLID: A large-scale semi-supervised dataset for offensive language identification. In **Findings of ACL/IJCNLP**, pp. 915–928, 2021.

[18] Perspective API. https://perspectiveapi.com/. Accessed: 2023-04-06.

[19] John Pavlopoulos, et al. Toxicity detection: Does context really matter? **arXiv preprint 2006.00998**, 2020.

[20] Yizhong Wang, et al. Self-instruct: Aligning language model with self generated instructions. **arXiv preprint 2212.10560**, 2022.

[21] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. **arXiv preprint 2108.13487**, 2021.

[22] Thomas Hartvigsen, et al. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In **ACL**, pp. 3309–3326, 2022.

[23] Paul Röttger, et al. Hatecheck: Functional tests for hate speech detection models. In **ACL/IJCNLP**, pp. 41–58, 2021.

[24] Baolin Peng, et al. Instruction tuning with gpt-4. **arXiv preprint 2304.03277**, 2023.

[25] 吉川克正 牧田光晴 中町礼文 佐藤京也 浅原正幸 佐藤敏紀小林滉河. 日本語有害表現スキーマの提案と評価. 言語処理学会 第 29 回年次大会 発表論文集, pp. 933–938, 2023.

[26] tiiuae/falcon-7b. https://huggingface.co/tiiuae/falcon-7b. Accessed: 2024-01-08.

[27] DeepL. https://www.deepl.com/ja/translator. Accessed: 2023-12-06.

[28] togethercomputer/redpajama-incite-7b-chat. https://huggingface.co/togethercomputer/RedPajama-INCITE-7B-Chat. Accessed: 2024-01-08.

[29] llm-attacks. https://github.com/llm-attacks/llm-attacks/tree/main/data/advbench. Accessed: 2024-01-011.

A 従来研究における不適切発話の定義・基準

従来研究で提案されている不適切発話の定義およびその基準についての詳細を述べる．[8] での不適切発話の定義・基準の詳細を表 2 に示す．[8] では，性的なコンテンツの中でも医療や教育の文脈による文章や，ヘイトスピーチの中でもヘイトスピーチに言及する中立的発言等は不適切として扱わずにデータセットを構築し判定モデルを学習している．

また，[9] での評価の尺度を表 3 に示す．

表 2: 不適切発話の定義・基準 [8]

lightgray **定義・基準**	**サブカテゴリ**	**適切**
S: 性的な内容 明示的または 暗示的な性的描写	S3：未成年者の性的内容 S2：違法となりうる性行内容 S1：違法行為でない性的内容 S0：医療・教育目的の性的内容	× × × ○
H: 憎悪的なコンテンツ 特定のアイデンティティ を理由とした，脅迫的， 侮辱的，軽蔑的なコンテンツ	H2：暴力や脅迫を呼びかける内容 H1：軽蔑的なステレオタイプへの支持 H0.a：ヘイトに言及する中立的な発言 H0.b：解説のための発言の引用	× × ○ ○
V: 暴力 肉体的な暴力を描写したり， その支持を示す内容	V2：極めて生々しい暴力 V1：脅迫，または暴力に対する支持 V0：文脈を考慮した中立的な内容	× × ○
SH: 自傷行為 生々しい自傷描写， または自傷行為への 支持を示すもの	なし	×
HR: ハラスメント 特定個人に迷惑をかけたり， 攻撃に利用される 可能性のある内容	なし	×

表 3: Perspective API における評価尺度 [18]

lightgray **定義**	**基準**
有害性 (TOXICITY)	失礼な発言，無礼な発言， 理不尽な発言で，会話から 離脱させる可能性が高いもの
重度有害性 **(SEVERE-TOXICITY)**	非常に憎悪的，攻撃的， 無礼なコメント，または ユーザが会話から離れる ようなコメント
アイデンティティへの攻撃 **(IDENTITY-ATTACK)**	アイデンティティを理由に， 誰かを標的にした否定的 または憎悪的なコメント．
侮辱 (INSULT)	人または集団に対する侮辱的， 扇動的または否定的なコメント
冒涜 (PROFANITY)	悪口，不敬または猥褻な表現， 冒涜的なコメント
脅迫 (THREAT)	個人または集団に対して， 苦痛，傷害，または暴力を 与える意図の記述が含まれる コメント
性的な内容 **(SEXUALLY-EXPLICIT)**	性行為，身体の部位，または その他の淫らな内容への言及 が含まれているコメント