制約が異なる指示で生成された文章に対する LLM 生成検出の頑健性

小池隆斗 [1]　金子正弘 [2,1]　岡崎直観 [1]
[1] 東京工業大学　[2]MBZUAI
{ryuto.koike@nlp., okazaki@}c.titech.ac.jp　masahiro.kaneko@mbzuai.ac.ae

概要

大規模言語モデル (Large Language Model; LLM) は，剽窃や誤情報の流布など様々な場面での悪用が懸念されており，LLM の生成文と人間が書いた文を頑健に分類する LLM 生成検出の実現は急務である．LLM の出力は生成指示に大きく依存するが，どのような指示で LLM にテキストを生成させると検出性能が変動するかは検証されていない．本研究では，文体や構造などの項目に着目し，指示による LLM の制約項目を変えることで検出器の頑健性の評価を行う．提案手法はサンプリングや言い換えによる頑健性評価と比較して，検出性能の標準偏差を最大 14.4 ポイントに増加させた．

1　はじめに

LLM は様々な指示に従って人間と同程度に尤もらしい文を生成することが可能となった [1, 2]．一方でそのような LLM の優れた生成能力を利用して，学生が宿題の回答に LLM の返答をそのままコピペしてしまったり，フェイクニュースを生成するといった LLM の悪用の懸念も明るみになっている [3, 4]．このような LLM の悪用を防ぐため，LLM による生成文と人間が書いた文を分類する LLM 検出器が最近多く提案され，一定の有効性が示されている [5, 6, 7, 8, 9]．

ユーザが LLM に指示を与えて文を生成する場面に注目すると，その指示文に含める制約（例: 生成文のフォーマット，文長など）はユーザごとに異なる [10]．そのような指示文の書き方の違いは，LLM が生成する文の質に加え，様々な下流 NLP タスク性能にも大きな影響を与えることが報告されている [11, 12, 13]．一方で，LLM 検出に関する研究は検出対象の文そのものに着目（例えば，LLM 文の言語的特徴の調査 [14, 15, 7, 16]）することが多く，その文がどのように生成されたかについては，注目されてこなかった．さらに，多くの LLM 検出ベンチマークでは，この文生成時の指示文の多様性は考慮されず，一つの決まった書き方の指示文で生成文が構成される [15, 7, 16]．これらを踏まえると，LLM 文生成時の指示文の書き方は生成文の検出難易度に影響を与えるか？という疑問が生じる．

本研究では，LLM 文生成時の指示文に含まれる制約がその生成文に対する検出難易度に大きな影響を与えることを報告する．LLM 検出の需要があるドメインの 1 つとして学生によるエッセイ執筆を想定し，Koike ら [8] が作成した問題文に対して LLM を使いエッセイを生成する．LLM に与える指示文中の制約として，Ke ら [17] が定義したエッセイの質に寄与する各項目に対応させ，制約文を人手で作成する（表 1）．そして，指示文中の制約が LLM 検出に与える影響として，各制約文を加えた指示文による各 LLM 文セットに対し，検出性能の標準偏差を測定する（図 1 右）．この時，検出性能への影響を与えうる他要素である，LLM 文の複数回生成（図 1 左）と指示文の複数回言い換え（図 1 中央）による検出性能の標準偏差との比較を行う．

評価実験の結果，指示文の言い換えや LLM 文を複数回生成する場合に比べ，指示文中の制約が検出性能の標準偏差を著しく増加させ（F1 値で最大 14.4 ポイント），その影響は大きいことがわかった．また制約を含む指示文から生成された LLM 文に対する分析から，指示文中の制約に従うという LLM の高い能力が，LLM 検出性能へ大きな影響力を与えることが示唆された．

2　提案手法

本節では LLM 文生成時の指示文中に含まれる制約が検出難易度に与える影響の検証方法について説明する．

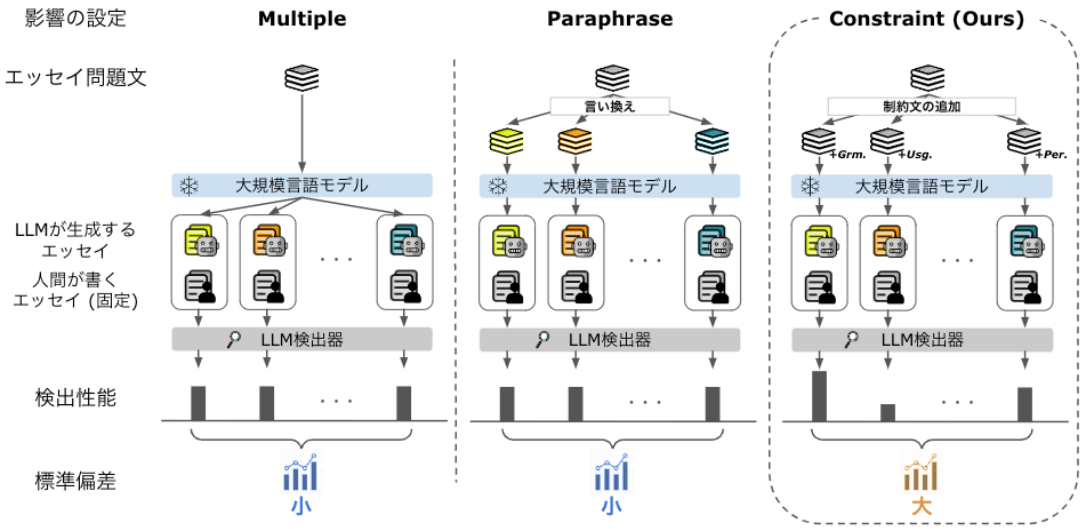

図 1　指示文中の制約による LLM 検出性能への影響の検証．異なる制約を含む指示文によって生成された各 LLM 文セットに対する検出性能の標準偏差を測定する（Constraint）．そして，LLM 検出性能へ影響を与えるその他の要素である，LLM 文の複数回生成（Multiple）と指示文の言い換え（Paraphrase）によって生じる検出性能の標準偏差との比較を行う．図中の Constraint の設定における Grm. や Usg. は表 1 のエッセイの質に寄与する各項目の略語である．

項目		制約文
Grammatically	(Grm.)	Your essay must be free of grammatical errors.
Usage	(Usg.)	Your essay must utilize a professional-level vocabulary.
Mechanics	(Mec.)	Your essay must be free of spelling and capitalization errors.
Style	(Sty.)	Your essay must include diverse word choices and sentence structures.
Relevance	(Rel.)	Your essay must follow the prompt.
Organization	(Org.)	Your essay must be logically organized.
Development	(Dev.)	Your essay must include concrete evidence that supports your opinion.
Cohesion	(Chs.)	Your essay must have a valid connection between paragraphs.
Coherence	(Chr.)	Your essay must have an effective transition throughout all paragraphs.
Thesis Clarity	(TC.)	Your essay must have a clear position through your essay.
Persuasiveness	(Per.)	Your essay must be persuasive to readers.

表 1　エッセイの質に寄与する各項目に対する制約文．

2.1　タスク設定

本研究で対象とするタスクは LLM 検出であり，具体的にはある 1 つのエッセイに対して LLM が生成したものか，人間が書いたものかを判定する．LLM 検出性能の測定のためには，一般に人間が書いた文群と LLM が生成した文群を混ぜ合わせたものをテストデータに用いる．我々は Koike ら [8] が作成したデータセットの一部である，エッセイ問題文と学生によるエッセイのペア群を利用する．そして，そのエッセイ問題文を元に LLM で生成したエッセイと学生によるエッセイを混ぜ合わせたものをテストデータとする．

2.2　制約による LLM 検出への影響の検証

図 2 に，LLM 文生成時の指示文中の制約が与える LLM 検出への影響の検証方法についての概要図を示す．指示文中に加える制約文について，Ke ら [17] が定義したエッセイの質に寄与する各項目に対する制約文を人手で作成する（表 1）．そして指示文中にそれぞれ異なる制約文を追加し，LLM を用いてエッセイを生成する．以下のように制約文は指示文中に挿入される．

```
Given the following problem statement,
please write an essay in [n] words.
[constraint]
Problem statement:
[problem_statement]
Essay:
```

ここで [n] はエッセイ問題文に紐づく人間エッセイ文の単語数であり，[constraint] は制約文，[problem_statement] はエッセイ問題文である．次に，異なる制約文を含む指示文から生成された各 LLM エッセイ文群と，事前に用意された学生によ

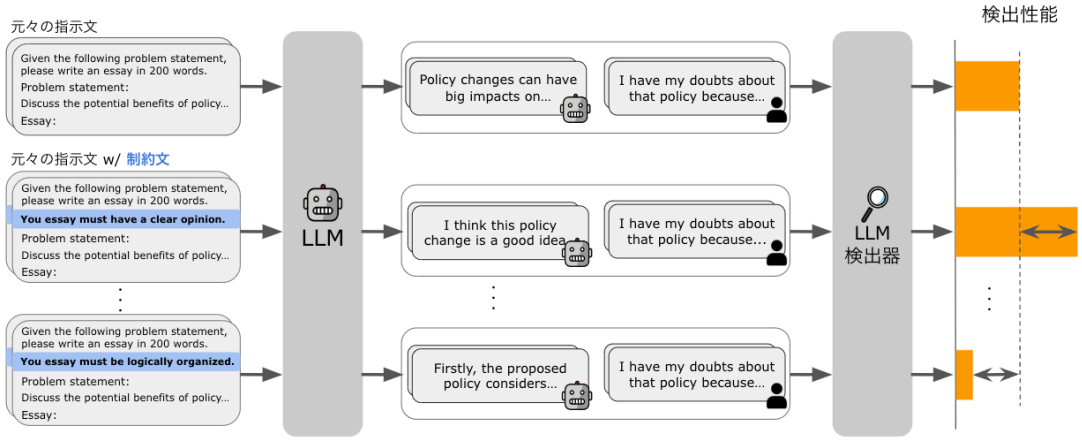

図 2　LLM 文生成時の指示文中の制約が与える LLM 検出への影響．異なる制約文を含む指示文から生成された LLM 文はその検出難易度に大きな変動が生じる．

るエッセイと混ぜ合わせ，LLM 検出器による分類を行う．最後に，それらの検出性能の標準偏差を制約による LLM 検出への影響と定義する．

3　実験

3.1　実験設定

エッセイ生成モデル　指示を与えてエッセイを生成させるモデルには，広く使われている LLM である ChatGPT（gpt-3.5-turbo-0613）と GPT-4（gpt-4-0613）を利用した．エッセイ生成時の温度パラメータは ChatGPT には 1.3, GPT-4 には 1.0 を適用した．

評価指標・データセット　評価実験にて対象とする全て LLM 検出器は 1 つの文に対して 2 値ラベルを生成する．そのため，LLM 検出の評価指標として 2 値分類タスクで一般に用いられている F1 値を採用する．評価セットとして，Koike ら [8] が作成したエッセイ問題文と学生が書いたエッセイのペア群を利用する．エッセイ問題文に対して LLM が生成したエッセイ 500 文と学生が書いたエッセイ 500 文をランダムシャッフルしたデータセットに対して LLM 検出器の性能を測定した．

LLM 検出器　本実験で対象とする LLM 検出器は，学習済み分類器として HC3 ChatGPT detector[1] と ArguGPT[2]，加えて In-Context Learning（ICL）[18] を用いた検出手法 [3] である．HC3 detector は，様々なドメインにおける ChatGPT による生成文を検出するために作成された HC3 データセットを用いて学習された RoBERTa ベースの検出器である [7]．ArguGPT は，LLM が生成したエッセイを検出するために学習された RoBERTa ベースの検出器である [16]．上記の検出器の学習データセットは，どちらも LLM 生成時の指示文の多様性を考慮せずに作成されている．ICL 検出手法は，学習セットから検索した人間文と LLM 文に正解ラベルを付けて，検出器としての LLM の few-shot 事例とし検出を行う．Koike ら [8] に倣った ICL 検出手法における詳細は，付録 A に記載する．

LLM 検出に影響を与える要素　評価実験では，指示文中の制約 (Constraint) 以外に LLM 検出に影響を与える要素として，LLM 文の複数回生成（Multiple）や指示文の複数回言い換え（Paraphrase）を比較対象とする．Constraint 設定では，指示文に制約を加えない場合と指示文に 11 個の各制約文を加えた場合の計 12 個の場合において LLM でエッセイを生成し，12 個の LLM 文セットを作成する．Multiple 設定では，制約を含まない各指示文からそれぞれ 12 個の LLM 文を生成[4] することで，12 個の LLM 文セットを作成する．Paraphrase 設定では，指示文に対して LLM で 12 個の言い換えを生成[5] する．次に，それぞれの言い換えられた指示文によって LLM にエッセイを生成させることで，12 個の LLM 文セットを作成する．指示文の言い換えに関する詳細は付録 A に記載する．

1）https://huggingface.co/Hello-SimpleAI/chatgpt-detector-roberta
2）https://huggingface.co/SJTU-CL/RoBERTa-large-ArguGPT
3）https://github.com/ryuryukke/OUTFOX
4）OpenAI Chat Completion API における n パラメータを使用：https://platform.openai.com/docs/api-reference/chat/create#chat-create-n
5）言い換えモデルには，ChatGPT（gpt-3.5-turbo-0613）を利用

エッセイ生成器	影響	LLM 検出器		
		HC3	ArguGPT	ICL
ChatGPT	Multiple	1.02	0.30	0.48
	Paraphrase	4.07	0.84	0.58
	Constraint	**12.76**	**6.69**	**1.15**
GPT-4	Multiple	1.09	1.14	0.68
	Paraphrase	3.42	2.43	0.69
	Constraint	**4.13**	**14.38**	**1.26**

表 2　3 つの設定（Multiple・Paraphrase・Constraint）において ChatGPT と GPT-4 それぞれが生成したエッセイに対する検出性能（F1 値）の標準偏差（%）の比較.

エッセイ生成器	影響	LLM 検出器		
		HC3	ArguGPT	ICL
Davinci-002	Multiple	1.07	0.15	0.78
	Paraphrase	**4.14**	**0.51**	**1.51**
	Constraint	1.44	0.32	1.17

表 3　3 つの設定（Multiple・Paraphrase・Constraint）において，Davinci-002 が生成したエッセイに対する検出性能（F1 値）の標準偏差（%）.

3.2　実験結果

評価実験では，指示文中の制約が LLM 検出に与える影響を検証するために，指示文から複数回 LLM 文を生成する設定（Multiple）と指示文を複数回言い換えてから LLM 文を生成する設定（Paraphrase）が引き起こす LLM 検出への影響との比較を行う．表 2 に，ChatGPT と GPT-4 がそれぞれの設定（Multiple・Paraphrase・Constraint）で生成したエッセイにおける検出性能の標準偏差の比較の結果を示す．エッセイ生成モデル，LLM 検出器の全ての組み合わせにおいて，Constraint 設定における検出性能の標準偏差が他の設定よりも大きく，最大で F1 値で 14.4 ポイントに達した．したがって指示文中の制約の違いは，LLM 文を複数回生成したり，指示文を言い換えするよりも LLM 検出に対して顕著に影響力を持つと示唆される．また，その構築時に指示文の多様性が考慮されていないベンチマークデータセットで学習された HC3 detector と ArguGPT は，どちらも指示文中の制約の違いによって検出性能が大きく変動していることがわかる．一方で，その手法の特性上，幅広い表現を事例として考慮しうる ICL では，指示文中の制約の違いによる生成文の表現の違いに対応でき，その検出性能の変動が比較的小さくなったと考えられる．

4　分析

本節では制約による LLM 検出への影響は，エッセイ生成器である LLM が持つ指示文中の制約に従う能力に起因するという仮説をおき，議論する．表 3 に，明示的に指示文に従うように学習されていない Davinci-002[6]（GPT-3）を用いてエッセイを生成させ，3 つの設定における検出性能の標準偏差を測定した結果を示す．

エッセイ生成器	制約に従ったかの一致率 (%) ↑
ChatGPT・GPT-4	88.28%
Davinci-002	49.29%

表 4　制約を含む指示文から ChatGPT・GPT-4, Davinci-002 が生成した文は実際にその制約に従っているか.

Davinci-002 がエッセイ生成器の場合，指示文に制約を加えた場合の検出性能の標準偏差は小さく，制約が持つ LLM 検出への影響は小さいことがわかる．この原因として Davinci-002 は ChatGPT や GPT-4 と比べ指示文中の制約に従う能力が低く，生成文の変化が小さいことが予想される．この指示文中の制約に従う能力をモデルごとに確認するため，近年様々な NLP タスクにおける評価器として一定の有効性が認められている [19] GPT-4 を用いた検証を行った．具体的には，各項目の制約を含む指示文による LLM 文 45 文ずつ計 495 文に対して，エッセイ問題文は同様だが制約を含まない指示文による LLM 文との比較を行い，よりその項目に従った文かを GPT-4 に判定させた．ChatGPT と GPT-4 についてはそれぞれが生成した文からランダムに選択した．

表 4 にその結果を示す．ChatGPT・GPT-4 における 88%の一致率に比べ，Davinci-002 においてはランダムに選択した場合の 50%と同程度の値となった．このことから，ChatGPT や GPT-4 は Davinci-002 に比べ指示文中の制約により従った文生成が可能であり，それが検出性能への大きな変動につながったことが示唆される．

5　おわりに

本研究では LLM 文生成時の指示文中に含まれる制約の違いによって，その生成文の検出難易度が大きく変動することを報告した．また分析の結果，文生成時の LLM による制約に従う能力が高いほど，指示文中の制約の違いによる検出性能のより大きな変動につながることが示唆された．今後ますます性能向上が期待される LLM においては，制約が持つ LLM 検出への影響はより大きくなりうる．そのため，指示文中の制約の違いによって生じる生成文の変化にも頑健な LLM 検出手法が求められる．

6）http://tinyurl.com/base-models-davinci

謝辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究（22501）により得られたものです.

参考文献

[1] OpenAI. Introducing ChatGPT, 2023. Accessed on 2023-05-10.

[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, 2023.

[3] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts, 2023.

[4] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. A survey on llm-generated text detection: Necessity, methods, and future directions, 2023.

[5] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models, 2023.

[6] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, 2023.

[7] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

[8] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. OUTFOX: LLM-generated Essay Detection through In-context Learning with Adversarially Generated Examples, 2023.

[9] Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, 2023.

[10] OpenAI. Prompt engineering guide, 2023. Accessed: 2023-10-10.

[11] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know?, 2020.

[12] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study, 2023.

[13] Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. Sentence simplification via large language models, 2023.

[14] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text, 2023.

[15] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. Deepfake text detection in the wild, 2023.

[16] Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models, 2023.

[17] Zixuan Ke and Vincent Ng. Automated Essay Scoring: A Survey of the State of the Art. In Sarit Kraus, editor, **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019**, pp. 6300–6308. ijcai.org, 2019.

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[19] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

Please compose a [n]-word essay based on the provided problem statement.
I kindly request you to compose an essay that adheres to the given problem statement, ensuring that it contains [n] words.
Could you kindly compose an essay containing [n] words based on the provided problem statement?
Please compose an essay of [n] words based on the given prompt.
Please compose an essay with a word count of [n], based on the provided problem statement.
Please compose an essay consisting of [n] words based on the provided problem statement.
I kindly request you to compose an essay with [n] words, based on the subsequent problem statement.
I kindly request you to compose an [n]-word essay based on the aforementioned problem statement.
Please compose an essay of [n] words based on the provided problem statement.
I am requesting an essay to be written in [n] words using the provided problem statement.
Please compose an essay in which you discuss the given problem statement, utilizing [n] to express your thoughts.
I kindly request you to compose an essay consisting of [n] words, using the problem statement provided below.

表 5　Paraphrase 設定において言い換えられた 12 個の指示文.

A　実験設定の詳細

ICL 検出手法　本研究では Koike ら [8] に倣い，ICL 手法では検出器に ChatGPT (gpt-3.5-turbo-0613) を利用する．そして，検出対象の文に紐づくエッセイ問題文と意味的に近いエッセイ問題文から生成された 5 つの LLM エッセイ文と人間エッセイ文のペアを学習セットから検索し，few-shot 事例とする．ここで，検出対象のエッセイ生成モデルの種類に関係なく，ChatGPT (gpt-3.5-turbo-0613) によるエッセイ群から few-shot 事例を取得した．また学習セットは，Koike ら [8] が作成した 14400 個のエッセイ問題文，人間エッセイ文，LLM エッセイ文の 3 つ組から構成される．

Paraphrase 設定での指示文の言い換え　以下に，Paraphrase 設定において言い換えられた後の指示文を示す．

```
[paraphrased_instruction]
Problem statement:
[problem_statement]
Essay:
```

ここで `[paraphrased_instruction]` は，元々の指示文中の “Given the following problem statement, please write an essay in [n] words.” を言い換えた文である．表 5 に，本研究の Paraphrase 設定における言い換えられた 12 個の `[paraphrased_instruction]` を示す．