対照学習を用いた hallucination 検出手法

山田美優 [1]　荒瀬由紀 [1]
[1] 東京科学大学
yamada.m.ee1b@m.isct.ac.jp, arase@c.titech.ac.jp

概要

大規模言語モデル (LLM) は様々な生成タスクで用いられ、高い性能を示すが、一方で hallucination と呼ばれる入力文章と矛盾する情報や入力からは事実性が検証不可能な情報を出力することが問題になっている。様々な hallucination 検出手法が提案されているが、その精度はまだ十分でなく、さらなる改善が求められている。本稿では、hallucination を含む出力が入力文章にない情報や推定困難な情報を含む点に着目する。この特性から入力文章と hallucination を含む出力ではそれぞれの埋め込み表現が乖離すると仮定し、対照学習を用いて hallucination 検出器を訓練する手法を提案する。実験の結果、提案手法はベースラインよりも hallucination 検出精度を大幅に向上させることが示された。

1　はじめに

大規模言語モデル (LLM) は現在様々な生成タスクで用いられ、高い性能を示している。しかし、その生成には hallucination と呼ばれる入力に記述された事実と異なる情報や入力から検証が難しいような情報が含まれることが多くある [1]。LLM の出力は流暢かつ自信を持った回答であることが多く、ユーザーが正しい知識を持っていない場合は hallucination を見抜くことが困難である。その結果、ユーザーは hallucination を含む情報を真実として受け入れてしまう危険性があり、その誤った情報が拡散されてしまう恐れがある。そのため、一般ユーザーが安心して LLM を使用することができておらず、LLM の普及や社会実装を妨げてしまっている。LLM がより多くの人々に使用されるためには、hallucination を高い精度で検出することが必要である。

hallucination 検出においては、モデルの内部状態を観測して検出する方法 [2] や、生成された文章と関連する文書を RAG (Retrieval Augmented Gemeration) によって選出し、入力文章としてモデルに与える方法が提案されている [3]。しかし、内部状態を観測して検出する方法はオープンなモデルにしか適用できない。また、RAG による参考知識を提供して検出する方法は検出精度がまだ十分でなく、更なる改善が求められている。さらに一口に hallucination といっても、入力文章と異なる事実を記述するもの (人名や数値の間違いなど) や、間違っているとは断定できないもの (検証不可能な情報や主観的な意見など) など様々なタイプが存在し、ある文章が hallucination を含むかどうかを判定するだけでなく、その文章で hallucination が起こっている箇所とタイプを特定して検出する手法もある [3][4]。その他にも、LLM の性能の向上に伴って、LLM にセルフチェックをさせることで hallucination を検出する手法 [5] や、LLM の出力を複数用意して比較することで hallucination を検出する手法 [6] も提案されている。

本研究では LLM の出力に hallucination が含まれるか否かを判定する問題に取り組む。先述の通り様々な hallucination 検出手法が提案されているが、既存研究では文章の埋め込みには着目していなかった。hallucination を含む出力は入力文章と矛盾する情報や入力文章からは推論できない情報を含むため、hallucination を含まない出力や入力文章とは乖離した埋め込み表現を持つと本研究では仮定する。逆に、hallucination を含まない出力と入力文章は埋め込み表現が近い可能性が考えられる。したがって、埋め込み表現に着目することで hallucination 検出精度を向上できると期待される。

本仮説に基づき、対照学習を用いることで hallucination 検出器を訓練する手法を提案する。今回は対照学習のために triplet loss を用いてモデルを学習する。triplet loss は、正例同士の embedding の距離が近く、負例と正例の embedding の距離が遠くなるように学習するための損失関数であり、主に画像認識の分野での対照学習において有効である

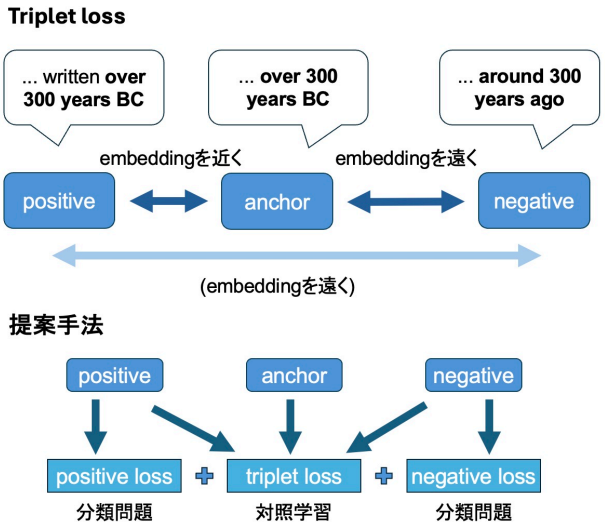

図 1　提案手法の概要

ことが知られている [7]。今回はテキストにおいてもこの手法が有効なのではないかと考え対照学習を行い、入力文章と hallucination を含まない文章の embedding の距離を近づけ、入力文章と hallucination を含む文章の embedding の距離を遠ざけることで hallucination 検出精度を向上させることを試みる。

実験の結果、QA タスクとニュース要約タスクにおいて triplet loss を用いた場合に hallucination 検出精度が大きく向上することが確認された。なお、本稿の実験に用いたコードは以下のレポジトリにて公開されている。 (https://github.com/miyu-y/nlp2025)

2　提案手法

本稿では与えられた文章内に hallucination が含まれるかを二値分類で判定する。対照学習に用いる triplet loss は図 1 上部に示す通り 3 つの入力 (anchor, positive, negative) を持ち、以下の式で計算される。

$$\texttt{triplet\_loss}(A,P,N) = \max(0, \alpha + d(A,P) - d(A,N)) \tag{1}$$

ここで、$d(A,P)$ は anchor と positive の距離、$d(A,N)$ は anchor と negative の距離、α はマージンである。本稿の hallucination 検出タスクでは、anchor は入力文章、positive は hallucination を含まない出力、negative は hallucination を含む出力とする。文章同士の距離の計算にはコサイン類似度を用いた。

$$d(x,y) = 1 - \texttt{cos\_sim}(x,y) \tag{2}$$

提案手法では、positive と negative の 2 つの文章について、hallucination を含むかを判定する分類問題を同時に学習する (図 1 下部)。すなわち、入力文章と positive および negative サンプルをそれぞれ結合して入力し、分類を行う。検出器全体の損失関数を以下に示す。

$$\begin{aligned}\texttt{loss}(A,P,N) = \texttt{triplet\_loss}(A,P,N) \\ + 0.5 \cdot \texttt{positive\_loss}(A+P) \\ + 0.5 \cdot \texttt{negative\_loss}(A+N)\end{aligned} \tag{3}$$

ここで、$\texttt{positive\_loss}(A+P)$ は入力文章と positive サンプル（hallucination を含まない出力）を結合して入力した場合の hallucination 判定に対するクロスエントロピー損失、$\texttt{negative\_loss}(A+N)$ は入力文章と negative サンプル（hallucination を含む出力）を結合して入力した場合のクロスエントロピー損失である。推論時は入力文章と LLM による出力を結合して検出器に入力し、hallucination を含むか否かを判定する。

3　評価実験

triplet loss を用いた対照学習によって訓練された hallucination 検出器で、hallucination 検出の精度が向上するかを検証するための実験を行う。

3.1　データセット

実験では RAGTruth[4] というデータセットを用いた。このデータセットは 1 つの入力文章に対してそれぞれ異なる 6 つの LLM (GPT-3.5-turbo-0613、GPT-4-0613[8]、Mistral-7b-Instruct[9]、Llama-2-7B-chat、Llama-2-13B-chat、Llama-2-70B-chat[10]) が生成した文章が含まれており、hallucination 箇所のラベルが付与されている。各ラベルには後述する hallucination のタイプも付与されている。

このデータセットは QA、Data-to-text、ニュース要約の 3 つのタスクから構成されており、hallucination を含む LLM 出力の割合がそれぞれ異なっている。QA タスクの入力文章は MS MARCO[11] に含まれる passage と question、出力は answer である。Data-to-text タスクの入力文章は Yelp Open Dataset[12] に含まれる、レストランについての json 形式に構造化されたデータと利用者のレビュー、出力はそれらの情報を自然文で文章化したものである。ニュース要約タスクの入力文章は CNN/Dairy Mail dataset[13] などに含まれるニュース記事、出力はその記事の要約である。

データセット内の hallucination は以下の 4 タイプに分類されている（表 1)。

表 1 hallucination タイプの統計

	Evi Confli	Sub Confli	Evi Base	Sub Base	All
QA	423	49	1562	893	2927
D2t	4174	63	3618	1435	9290
要約	727	89	1057	199	2072
合計	5324	201	5237	2527	14289

表 2 データセットの分割 (() 内は三つ組の数)

	訓練データ	検証データ	テストデータ
QA	4614 (1160)	420 (100)	900 (150)
Data-to-text	4878 (1424)	420 (127)	900 (291)
ニュース要約	4338 (1176)	420 (110)	900 (192)
合計	13830 (3760)	1260 (337)	2700 (633)

- Evident Conflict : 数値や名前の間違いなど、入力と明らかに異なる情報
- Subtle Conflict : 入力が意図した意味とは異なる情報の提供、入力とニュアンスが異なる記述など
- Evident Introduction of Baseless Information : 入力に含まれる情報では裏付けられない情報など
- Subtle Introduction of Baseless Information : 検証不可能な情報や主観的な意見など

入力文章に対する 6 種類の LLM 出力から（入力文章、hallucination を含まない出力、hallucination を含む出力）の三つ組みを抽出した。[1)] 元のデータセットには 17790 個の出力が含まれていたが、上記の処理を行った結果 4725 個の組み合わせが作成された。RAGTruth には検証データが存在しなかったため、訓練データの中からランダムにサンプリングすることで検証データを作成した。それぞれのサンプル数は表 2 に示す。

3.2 比較手法

本実験では検出器のベースモデルとして RoBERTa-base[14] または phi-3.5-mini-instruct[15] を用いた。RoBERTa-base では学習率 5e-6、phi-3.5-mini-instruct では学習率 1e-6 でそれぞれ 10 エポックの学習を行った。提案手法における α は 1.0 とした。[2)] さらに以下の 2 つのベースラインと比較する。

入力文章を用いないベースライン (no_doc) 入力文章を用いず判別対象である出力文章のみを入力して、hallucination が含まれるか判別する hallucination 検出器を訓練した。

1) サンプル間で hallucination を含む出力、含まない出力が重複して出現しないよう組み合わせを作成した。
2) 予備実験の結果、検出性能は α の設定に大きく依存しないことが示されている。

表 3 実験結果 (hullucination 検出の F1 値)

LM	手法	QA	Data-to-text	要約	All
RoBERTa	no_doc	42.8	84.0	39.9	64.5
	with_doc	46.1	**84.6**	37.7	66.5
	triplet	**74.7**	82.6	**67.1**	**75.7**
phi-3.5	no_doc	50.8	83.5	28.7	67.1
	with_doc	62.6	**86.0**	46.9	73.7
	triplet	**89.9**	77.7	**70.0**	**78.9**

表 4 hallucination タイプ別の検出成功数と割合 (「全体」はテストセット中の hallucination タイプの数)

		Evi Confli	Sub Confli	Evi Base	Sub Base
QA	no_doc	7	0	93	31
		(23.3%)	(0.0%)	(59.6%)	(63.3%)
	triplet	27	0	141	47
		(90.0%)	(0.0%)	(90.4%)	(95.9%)
	全体	30	0	156	49
D2t	no_doc	428	3	392	76
		(87.5%)	(60.0%)	(84.3%)	(80.0%)
	triplet	417	3	369	67
		(85.2%)	(60.0%)	(79.4%)	(70.5%)
	全体	489	5	465	95
要約	no_doc	18	2	31	10
		(18.0%)	(18.2%)	(26.5%)	(62.5%)
	triplet	39	3	66	13
		(39.0%)	(27.3%)	(56.4%)	(81.3%)
	全体	100	11	117	16

入力文章を用いた分類を行う検出器 (with_doc) 入力文章と判別対象である出力文章を結合して入力し、hallucination が含まれるか判別する hallucination 検出器を訓練した。

4 実験結果と分析

表 3 に各ベースモデルを用いた際の hullucination 検出の F1 値を示す。QA と要約タスクでは phi-3.5-mini-instruct の性能が高いが、Data-to-Text タスクでは RoBERTa と phi-3.5-mini-instruct は同程度となっている。全体的に見て、triplet、with_doc、no_doc の順に精度が高い。triplet loss を用いた場合、特に QA と要約のタスクにおいて精度の大きな向上が見られた。一方で、Data-to-text のタスクにおいては精度が低下している。

表 4 に phi-3.5-mini-instruct を用いた際にテストセットにおいて検出できた事例の hallucination のタイプ毎の統計を示す。以降では、対照学習を行ったことで検出できるようになった hallucination、また検出できなくなった hallucination について分析を行う。比較は検出性能の高かった phi-3.5-mini-instruct における no_doc と triplet の結果を用いる。

表 5　Hallucination 検出に成功・失敗した例（下線部が hallucination）

要約タスクにおける成功例	
入力	A nuclear submarine being repaired at a Russian shipyard has caught on fire, ... “Nuclear fuel from the sub’s reactor has been unloaded,” he reportedly said. “There are no armaments or chemically active, dangerous substances, fissionable materials on it,” Gladyshev said to Tass.
出力	A nuclear submarine being repaired in Russia caught fire due to accidental welding damage, causing an explosion that released radiation into the air. ... The damaged sub is currently being dismantled and decommissioned.（放射能が放出されたという記述、損傷した潜水艦が廃止されたという記述は入力にないため Evident Baseless）
Data-to-Text タスクにおける失敗例	
入力	{“name”:“Finch & Fork”, “address”:“31 W Carrillo St”, “city”:“Santa Barbara”, ... “review_info”:[{“review_stars”:5.0 ...}...]}
出力	Finch & Fork is a breakfast and brunch restaurant located in the heart of Santa Barbara（Santa Barbara の「中心地」にあるとは記述されていないが否定できないため Subtle Baseless）, within a luxury hotel. ... Finch & Fork is a popular spot for both locals and visitors（入力に記述されていないため Evident Baseless）, and is especially known for its bottomless brunch deals. ...

4.1　QA・ニュース要約タスク

QA タスクおよびニュース要約タスクでは triplet loss を用いた場合に精度が大きく向上している。表 4 より、特に QA タスクではどの hallucination タイプにおいても 90%以上の精度で検出できるようになったことがわかる。中でも特に Evident Conflict タイプ (数値や名前の間違いなど) の hallucination について効果が大きいことがわかる。

表 4 より、ニュース要約タスクにおいては特に Evident Baseless タイプ (入力に含まれる情報では裏付けられない情報など) の hallucination について効果が大きいことがわかる。表 5 にニュース要約タスクにおいて実際に検出できるようになった Evident Baseless タイプの例を示す。また、Subtle Baseless タイプ (検証不可能な情報や主観的な意見など) の hallucination については検出精度が 80%を上回ることができた。一方で、QA タスクでは大きな精度の向上が見られた Evident Conflict タイプ (数値や名前の間違いなど) について、提案手法では no_doc に比較して検出精度は大きく改善したが、検出成功割合は 39.0% にとどまっている。また Subtle Conflict タイプ (入力が意図した意味とは異なる情報など) の hallucination については依然として検出精度が低いことがわかる。

4.2　Data-to-text タスク

Data-to-text のみ triplet loss を用いて対照学習を行った場合に精度が低下している。このタスクにおける入力文章は表 5 に示す通り構造化データであるのに対し、大規模言語モデルによる出力文章は自然文である。そのため構造化データと自然文の埋め込みを空間上で近づける・遠ざける学習を行う対照学習が悪影響を及ぼしている可能性が考えられる。表 4 より、特に Subtle Baseless タイプ (検証不可能な情報や主観的な意見など) と Evident Baseless タイプ (入力に含まれる情報では裏付けられない情報など) の hallucination 検出の性能が低下している。表 5 に実際に検出できなくなった Subtle Baseless タイプ、Evident Baseless タイプの例を示す。

5　おわりに

本稿では triplet loss を用いた対照学習によって hallucination 検出器を訓練する手法を提案した。実験結果より、triplet loss を用いた場合、QA とニュース要約のタスクにおいて hallucination 検出精度の大きな向上が見られた。一方で、Data-to-text のタスクにおいては精度が低下している。

本稿では hallucination の 4 タイプに基づいて分析を行ったが、タスクによって対照学習による検出精度の向上に効果があるタイプが異なることが明らかとなった。今後は、hallucination タイプごとの特性を考慮した検出手法を検討することで、検出精度の向上を図りたい。また構造化データのような自然文でない入力を適切に扱う手法についても検討する予定である。

参考文献

[1] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. arXiv preprint arXiv:2307.10169, 2023.

[2] Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On large language models' hallucination with regard to known facts. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024.

[3] Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. In First Conference on Language Modeling, 2024.

[4] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024.

[5] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. 2023.

[6] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. 2023.

[7] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal et al. Anadkat. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[9] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and Lucile et al. Saulnier. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and Shruti et al. Bhosale. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[11] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. 2016.

[12] Yelp. Yelp open dataset, 2017. http://www.pluto.ai.kyutech.ac.jp/NLP/.

[13] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer- generator networks. In In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[15] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.