

架空語に対する LLM の知ったかぶりの自動評価

岩本 蘭^{1,2} 金山 博¹ 村岡 雅康¹ 吉田 一星¹

¹ 日本アイ・ビー・エム株式会社 東京基礎研究所 ² 慶應義塾大学
 ran.iwamoto1@ibm.com {hkana, mmuraoka, issei}@jp.ibm.com

概要

LLM が起こす幻覚には様々な種類がある。例えば、存在しそうで存在しない「架空語」の意味を尋ねた際、LLM は「知らない」と答えるのが望ましいが、誤った説明や別の語の意味を返すことがある。本論文では LLM が幻覚を引き起こさずに適切に未知を示す能力を評価する方法を提案する。架空語を自動生成し、架空語について問う質問に対する LLM の応答を分類することでこの能力を評価する。

1 はじめに

大規模言語モデル (LLM) は人間に匹敵するテキスト生成能力 [8] をもつ一方、しばしば幻覚を引き起こす [5]。医療や法律など専門性を要する分野において、幻覚による誤情報が意思決定に悪影響を与える可能性が指摘されている [9]。特に学習データに存在しない知識について質問したときに LLM が幻覚を起こす傾向は顕著で [6]、Huang ら [3] は、LLM が自らの知識の限界を評価できないことが原因と指摘している。本研究では、LLM が未知の知識に直面した際に幻覚を起こす度合いを評価する。

LLM が「知らない」と応答する、応答の棄権に関する研究として、応答の特徴の分析 [11] や、棄権すべき質問の自己検知 [12] に関する研究が挙げられる。一方で、棄権能力を評価するためのデータセットの構築に焦点を当てた研究は限られる。例えば、Li ら [7] は人手でデータセットを作成したが、作成コストが高いという課題がある。

本研究では、LLM の棄権能力を評価する簡便かつ再現性の高い方法として、架空語データセットの活用を提案する。存在しない「架空語」を自動生成し LLM の応答を分析することで、モデルが自身の無知を適切に示す能力を評価する枠組みを提供する。その結果、LLM は架空語の意味を自分で推測して幻覚を引き起こすことがわかった。

本論文の主な貢献は以下の3つである。



図1 架空語の返答の3つの分類

- LLM の棄権能力を判別するための架空語を定義した
- 実在する語を用いて自動的に架空語候補を生成し、連想語のアイデアを用いて尤もらしい架空語を選別した
- LLM に架空語の定義を生成させ、その応答を分類して棄権能力を分析した

2 提案手法

架空語の定義を LLM に尋ねた際の返答の例を図1に示す。LLM の棄権能力を評価するため、架空語候補の生成と絞り込み、LLM の応答の自動分類を行う。

2.1 架空語の定義

ここでは架空語の定義と性質を述べる。架空語を、多くの人が使用しない語と非実在語を合わせたものの中で、普段使いそうな尤もらしい語と定義する。尤もらしさの定義は 3.1 節で述べる。架空語には、その定義から、既知の知識ベースや辞書で定義される意味が存在しないか利用可能でない。つまり、言語モデルが未知の単語を処理する方法を評価するデータとして架空語を活用できる。

実験的には、何かしらの方法で文字から単語を作り出し、その単語がある言語の日常会話をカバーする十分大きなコーパスに存在しないとき、その言語の架空語と呼ぶ。架空語を構成する文字は、その言語の文字規則に従う。例えば英語の架空語ではアルファベットが用いられ、日本語では漢字とひらがなとカタカナを使う。本研究では日本語の漢字2文字からなる架空語について検証する。今回はルールベースで架空語候補を作成する。

2.2 架空語候補の自動生成

本研究では、混合法と逆転法を用いて日本語の架空語候補を生成する。2つの手法とも、ある程度大きなコーパスに存在する語である実在語を用いる。混合法では、2つの実在語を選び、以下のようにその単語同士を文字単位で混ぜる。

- 来月 + 決裁 = 来裁

逆転法では実在語を文字単位で2つの部分に分けて、以下のように反転する。

- 料金 → 金料

本研究では日本語の BCCWJ コーパスにおける単語の出現頻度を数えた頻度表を用いて、頻度が0であるものを架空語候補とした。

これらの架空語候補には、実在しそうな尤もらしい語とそうでない語が存在する。LLM の幻覚の誘発には尤もらしい語を用いる方が適切に機能すると考えた。また、尤もらしさとは単語が現実世界でどの程度使われているかという期待値を表し、それは単語の出現頻度と関係すると考えた。尤もらしさの性質を持つ架空語を選定するために連想語を用いた。連想語を用いるアイデアは、Utsumi ら [10] の研究で、関連する単語はベクトル空間上で刺激語 (連想元の語) の近くに配置されるという知見に基づく。

そこで我々は、使用頻度が低くなる、すなわち尤もらしさが低くなるにしたがって、連想語と刺激語の cos 類似度が小さくなるという仮説を立てた。

2.3 連想語を用いた架空語の絞り込み

この節では、架空語候補から架空語を抽出するために連想語を用いるプロセスについて説明する。LLM¹⁾に連想語を求めるプロンプト1を入力する。

—— プロンプト 1 ——

w から連想される単語を2つのみ簡潔に教えてください。

w には実在語や架空語候補を代入する。そこで出てきた2つの連想語の cos 類似度を FastText [2] の分散表現を用いて測る。分散表現のデータには単語とベクトルの組が含まれ、連想語が分散表現にない場合は実在語や架空語候補を検証リストから外した。

LLM として Llama3 [1] を用いた。実在語として頻度表に載っている漢字二字熟語を抽出して用い

1) 評価対象ではなく架空語の生成にのみ用いる LLM である。

表1 Llama3 でのそれぞれ 500 語の単語の連想語の cos 類似度の平均値。高頻度語の cos 類似度が高い。

単語の種類	cos 類似度
実在語-高頻度	0.371
実在語-低頻度	0.360
架空語候補-混合法	0.316
架空語候補-逆転法	0.323

表2 実際の架空語候補の連想語と cos 類似度。定性的に尤もらしい単語の cos 類似度が高い。

架空語候補	連想語 1	連想語 2	cos 類似度
果具	果物	野菜	0.748
投額	寄付	寄贈	0.620
工声	アクロニム	BGM	0.073

た。日本語の頻度表として BCCWJ 語彙表²⁾を用いた。その中より高頻度語としての上位 1000 語以内のものを、低頻度語として上位 4000-5000 語以内のものをを用いた。

実在する高頻度語と低頻度語、架空語候補 (混合法、逆転法) のそれぞれ 500 語の連想語をモデルに出力させ、cos 類似度を平均したものを表1に示す。

高頻度語、低頻度語、架空語候補の順に cos 類似度が高くなっており、連想語の cos 類似度で単語の尤もらしさを測ることができると言える。

混合法で作成した架空語候補の例を表2に示す。cos 類似度が高い“果具”の果は実りという意味で、具は具材などの意味で用いられ、食べ物に関する類似した連想語が出てきやすい。“投額”は投資、投票などと同じ構造を持つため尤もらしく見える。一方、“工声”は漢字同士が意味的に補完し合っていないため、尤もらしくないと感じられる。これらの直感は、cos 類似度の大小と対応している。

3 実験: 応答自動分類

分析のために LLM の応答を自動分類する。

3.1 架空語に対する LLM の応答の分類

LLM に架空語に関するプロンプト2を与えた。

—— プロンプト 2 ——

w という言葉の定義について、 w とは～、から始める形で日本語で教えてください。

観察により、応答のほとんどが3つに分類されることがわかった。応答種類を以下に記す。

2) <https://clrd.ninjal.ac.jp/bccwj/freq-list.html>

表3 LLMの架空語に対する3種類の応答の例。曲解の例は「州杭」に対する返答。

応答種類	応答例
未知の認知	「桐味（きりあじ）」という言葉は、一般的な日本語の中では認識されていないため、特定の定義は存在しません。そのため、「桐味」という言葉は造語か、誤って用いられている可能性があります。仮に、「桐」という漢字が木材の「桐の木」を指し、「味」が味覚や風味を表すことから、この言葉を連想的に解釈すると「桐の木に関連した風味」や「桐の木材に由来する特徴」として考えられるかもしれませんが、これはあくまで推測に過ぎず、実際の意味としては確立されていません。
捏造	田角（たつのす）は、漢字を美しく見せるための余白のこと。
曲解	「杭州」という言葉の意味は、中国の地理的な場所を指し、東は揚子江の南岸、西は太湖の北岸、北は銭江の南岸、南は桐廬江の北岸にあります。

表4 プロンプト2に対するLLMの返答の分類。

モデル	cos 類似度	未知の認知	捏造	曲解
Llama3	high	0	194	6
Llama3	low	4	193	3
GPT-4o mini	high	2	198	0
GPT-4o mini	low	7	193	0

表5 プロンプト3に対するLLMの返答の分類。

モデル	cos 類似度	未知の認知	捏造	曲解
Llama3	high	173	27	0
Llama3	low	184	16	0
GPT-4o mini	high	200	0	0
GPT-4o mini	low	199	0	1

- 未知の認知: 知らないと答える。理想的な返答。
- 捏造: 架空語の意味を答える。幻覚。
- 曲解: 別の似た単語の意味を答える。幻覚。

例えば、LLMの応答に“存在しません”、“辞書に載っていません”、“造語”、などと入っている場合は未知の認知と判断した。応答例を表3に示す。

3.2 実験準備

分類の正確さと結果の分析の2つの実験を行う。2.2節と3.1節に示した方法で架空語を混合法と逆転法、cos 類似度が0.75以上と0.5以上0.75未満の2種類の範囲それぞれで400語ずつ作成した。モデルとしてLlama3 70B [1]、GPT-4o mini [8]を用いた。

3.3 分類精度

まず、LLMの返答を正確に自動分類できているかを評価する。cos 類似度が高い架空語に対してプロンプト2で返答を作成、生成された返答の分類結果を手動でアノテーションし、自動分類と比較した。文字列マッチングを用いた分類の誤り数はLlama3とGPT4o-mini 両方とも0だった。機械学習モデルを使わずに高精度で3種類に分類できている。

3.4 分類結果

プロンプト2と、架空語の存在を暗に仮定したプロンプト3を用いてモデルごとの分類結果を示す。

プロンプト3

wという単語が学習データセットに存在しない場合、そのことを明確にして意味を知らないと答えてください。含まれていればその意味を教えてください。ある単語が学習データセットに存在しない場合、あなたはその意味を知りません。必ず日本語で教えてください。

LLMの返答を自動分類した結果を表4,5に示す。架空語の存在を想定していないプロンプト2では理想的な解答である“未知の認知”と答えられないモデルが多く、モデルやプロンプトの改善の必要を示唆する。プロンプト2,3両方とも、連想語のcos 類似度が高く尤もらしい架空語ほど未知の認知と答えられるケースが少なかった。

4 追加実験: 実在語の応答自動分類

実在語(高頻度語)の言葉の定義についてもプロンプト2,3を用いてLLMに尋ねた。先ほどの実験と異なり、未知の認知ではなく正解の意味を答えるのが正しい返答である。分類判定は人手で行った。結果を表6に示す。Llama3では“奥手”や“弘化”(年号)などについて捏造に分類された回答があった。また“結城”(茨城県の地名)について、“水を利用して城を守る城の形態”という曲解が見られた。

プロンプトの差異による結果を比較すると、GPT-4o miniではプロンプト2より未知の認知を促

表6 実在語 200 単語に対する LLM の返答の分類結果。

モデル	Prompt	正解	未知の 認知	捏造	曲解
Llama3	定義 1	161	0	39	0
Llama3	定義 2	145	35	20	0
GPT-4o mini	定義 1	200	0	0	0
GPT-4o mini	定義 2	190	10	0	0

表7 200 単語に対する定義文からの例文選択。

	Llama3	GPT-4o mini
正解数	156	189

すプロンプト 3 の方が正解数が少し下がる。例えばプロンプト 3 では“好位”(競馬用語)の意味を知らないと答えていた。以前の実験と合わせて考察すると、プロンプト 3 は架空語に関して未知の認知能力を高めるが実在語の正解数が下がる。

5 幻覚抑制能力の分析

LLM が架空語にどのように対応するかを明らかにし、その挙動を通じて LLM の知識境界の理解を深めるために、架空語の意味を新たに定義して LLM に与える。LLM が未知の知識に直面した際に過剰な推測や誤った知識生成をするという幻覚を引き起こさず、与えられた定義や文脈のみに基づいて回答ができるかという幻覚抑制能力を評価する。架空語の意味と例文を手作業で多数作るのは困難な作業のため、実在語の定義文と例文を用いて、実在語の部分を架空語に置き換える。

日本語 WordNet[4] の漢字二文字以上の単語の語義の定義文と例文を使い、定義文が架空語の意味の唯一の手がかりになるような問題を作成する。プロンプトの例をプロンプト 4 に示す。この例の場合、正解は a である。

— プロンプト 4 —

a から d のの中から選択してください。単語“旋律”の意味は [例: 特定のピッチの音符を正確に生産する特性] である。その単語の正しい例文はどれですか。

a 彼は正しい旋律で歌うことができない。

b 夜明けで夜から旋律が変わった。

c 病気の旋律。

d それは旋律の印であった。

5.1 結果

結果を表 7 に示す。本実験の指示は「a-d から選択してください」のため、「その単語は知らない」という回答は得られなかった。正解の例文を選んだ際、モデルは与えられた情報のみで正解を導いた。不正解の問題のプロンプトとモデルの回答例も示す。

— プロンプト 5 —

a から d のの中から選択してください。単語「野粗」の意味は何かを、平易かまたは理解できるようにすることです。その単語の正しい例文はどれですか。

a: ドイツは第一次世界大戦の後で野粗を支払うことができなかった

b: それは、島の野粗環境を変えた

c: 私はその事故の野粗を聞いた

d: どんな野粗で、社会的な科学がありますか?

正解は c だが、GPT-4o mini は b と回答した。誤答の原因を分析するため、プロンプト 6 でモデルが架空語の意味をどう捉えているか調べ、回答 1 を得た。

— プロンプト 6 —

野粗という単語の意味を教えてください。

— 回答 1 —

「野粗」という言葉は、一般的な日本語としては使用されておらず、辞書にも載っていない可能性が高いです。ただし、以下のように構成要素から推測可能です。「野」: 野生、自然、手つかずの状態を表す。「粗」: 粗野、荒々しい、未整備の状態を指す。これらを組み合わせると、「野粗」は「荒々しく、未整備な自然」や「手つかずで素朴な自然」を表す比喩や形容として使われる可能性があります。

モデルが推測する意味では答えは b となる。つまり既存の知識が推測に依存して幻覚が発生している。

6 まとめ

本論文では LLM の知ったかぶりの能力を評価するために架空語を定義し、自動生成と応答自動分類手法を提案した。LLM の返答を未知の認知、捏造、曲解に分類した。また、LLM は架空語の意味を勝手に推測して回答する場合があることを示した。

参考文献

- [1] AI@Meta. Llama 3 model card. 2024.
- [2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
- [3] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2024.
- [4] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. Development of the Japanese WordNet. In *Proceedings of the LREC*, Marrakech, Morocco, 2008.
- [5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, Vol. 55, No. 12, 2023.
- [6] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*, 2023.
- [7] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the EMNLP*, pp. 6449–6464, 2023.
- [8] OpenAI and team. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.
- [9] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the CoNLL*, pp. 314–334, 2023.
- [10] Akira Utsumi and Daisuke Suzuki. Word vectors and two kinds of similarity. In *Proceedings of the COLING/ACL*, pp. 858–865, 2006.
- [11] Bingbing Wen, Bill Howe, and Lucy Lu Wang. Characterizing LLM abstention behavior in science QA with context perturbations. In *EMNLP*, pp. 3437–3450, 2024.
- [12] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Knowing what LLMs DO NOT know: A simple yet effective self-detection method. In *Proceedings of the NAACL-HLT*, pp. 7051–7063, 2024.

A 架空語リスト

使用した架空語の例を表 8 に示す。

表 8 混合法と逆転法で作成した架空語の例

手法	架空語	元の単語
混合法	平色	平凡&虹色
	筆病	筆跡&仮病
	安額	安心&金額
	総談	総称&談話
	社期	社会&期日
逆転法	金料	料金
	黙沈	沈黙
	婆産	産婆
	箱巢	巢箱
	身白	白身