

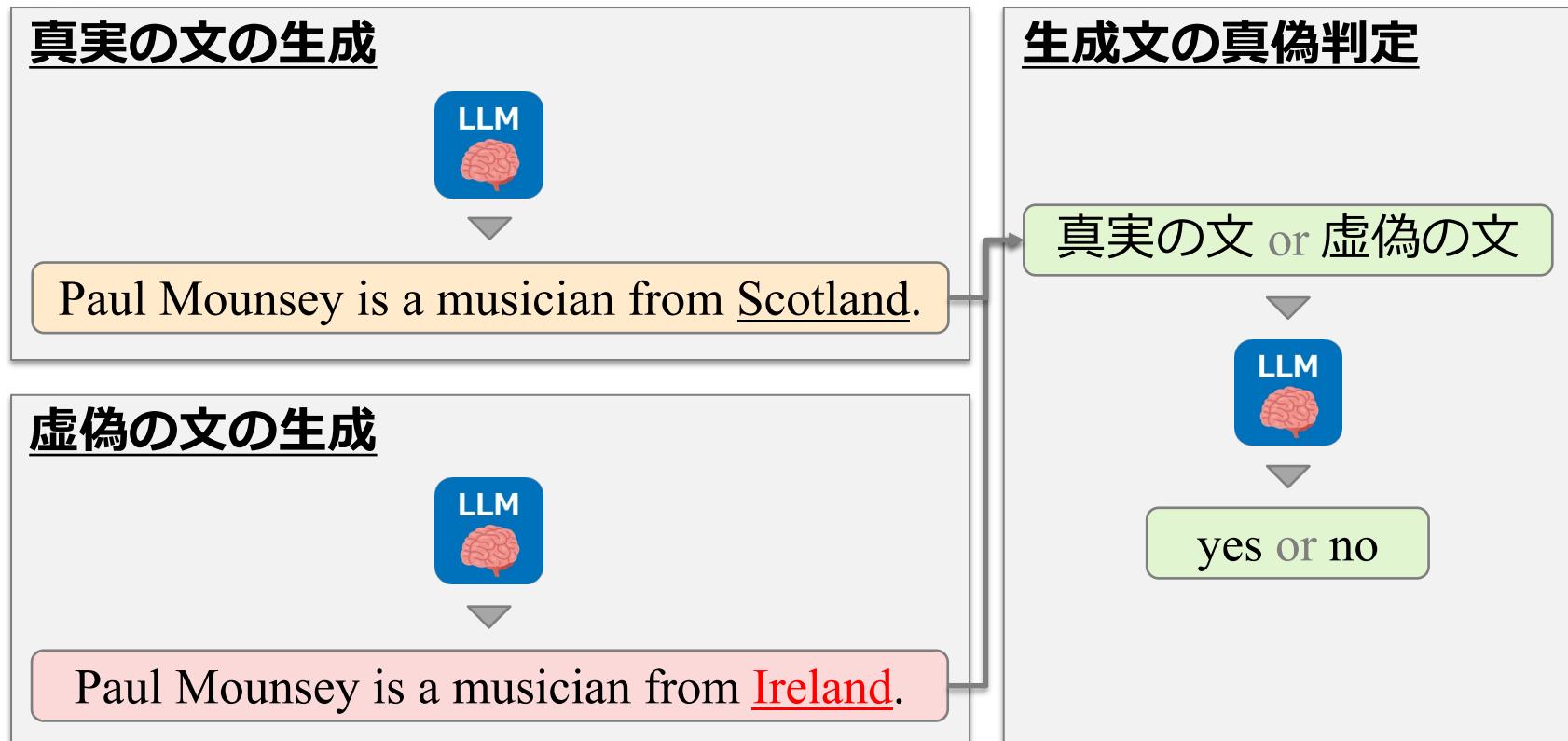
大規模言語モデルは自身の Hallucinationを検知できるか？

NTT人間情報研究所

○門谷宙, 西田光甫, 西田京介, 斎藤邦子

本研究の目的

- RQ : **LLMは自身のhallucinationを検知できるか？**
- hallucination検知をある文の真偽判定タスクとして定式化し、LLMのhallucination検知性能を測るフレームワークで検証



本研究の貢献

- LLMのパラメータに内包されている知識を引き出すために
Chain-of-Thought (CoT) [1] を用いる真偽判定手法を提案
- 実験の結果, GPT3.5-Turboは自身のhallucinationを58.2%検知できることを確認
- 詳細な分析により, hallucination検知率はLLMのパラメータに内包されている知識量に関係することが判明

目次

- 背景
- 検証フレームワーク (FW)
- 実験
- 分析

背景 : hallucination

- LLMが嘘をつき事実に基づかない応答をする現象
- アプリケーションの信頼性を低下させるため、開発者は検知して防ぐ必要がある
- 本研究の目的はhallucination検知

SO You
久保建英について教えてください。

ChatGPT
久保建英（くぼたけふさ）は、日本のサッカー選手であり、非常に注目されている若手選手の一人です。彼は2001年6月4日に神奈川県横浜市で生まれました。
川崎市

久保建英は非常に幼い頃からサッカーの才能を示し、幼少期からスペインの強豪クラブであるFCバルセロナのユースチームでプレーしました。その後、2019年にはレアル・マドリードと契約し、同クラブのトップチームに昇格しました。

その後、久保建英はレアル・マドリードからのレンタル移籍で複数のクラブに所属しました。スペインのラ・リーガでの経験を積むために、マジョルカやガラタサライ（トルコのクラブ）などでプレーしました。

背景：既存のhallucination検知手法

- 内部状態を用いる手法 [2]
 - トークン予測確率からLLMの確信度を推定
 - トークン予測確率が低い部分にはhallucinationが含まれている可能性が高いと仮定することで検知

 LLM単体で適用可能
 内部状態にアクセスできないLLMには使えない
- 外部知識を用いる手法 [3]
 - ファクトチェックの分野で活発に研究されてきた
 - データベースやウェブから根拠となる文書を取得し、生成テキストに矛盾する内容がないか確認することで検知

 内部状態にアクセスできないLLMにも使える
 外部モジュールが必要で、各アプリケーションに適した検索システムを構築することはコストが高く困難

[2] <https://arxiv.org/abs/2302.09664>

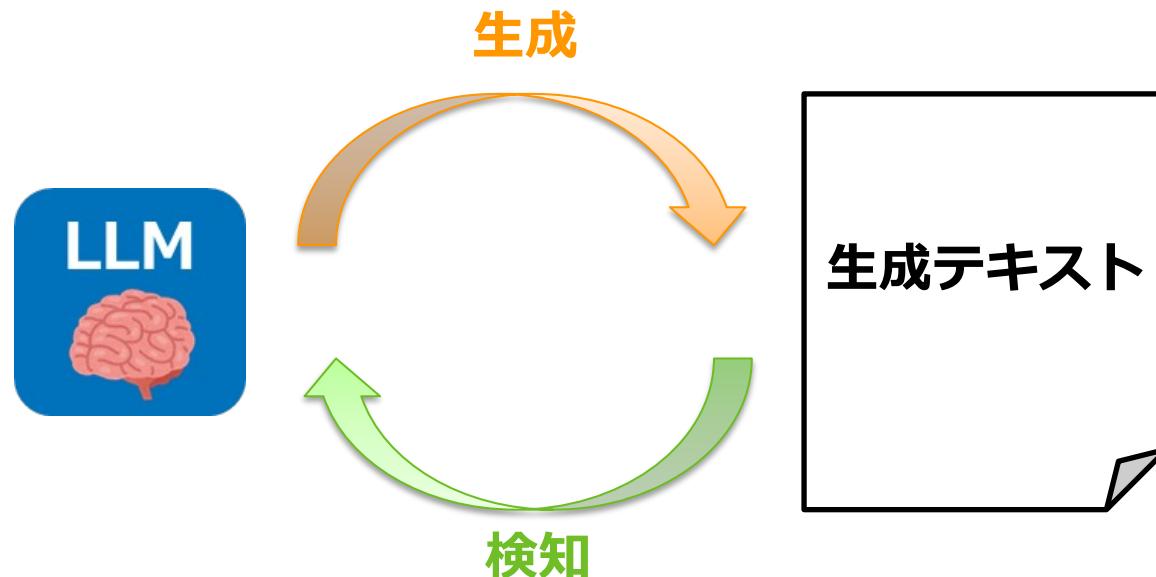
[3] <https://aclanthology.org/2022.tacl-1.11/>

背景：リサーチクエスチョン (RQ)

- LLMが自身で生成したhallucinationを検知することは困難だと考えられる
- LLMをhallucination検知器として用いることができれば
 - 😊 内部状態にアクセスできないLLMにも適用可能
 - 😊 外部モジュールが不要で、すべてのユーザが手軽に使える

背景：リサーチクエスチョン (RQ)

- LLMが自身で生成したhallucinationを検知することは困難だと考えられる
- LLMをhallucination検知器として用いることができれば
 - 😊 内部状態にアクセスできないLLMにも適用可能
 - 😊 外部モジュールが不要で、すべてのユーザが手軽に使える
- RQ : **LLMは自身のhallucinationを検知できるか？**
 - LLMのhallucination検知性能を測るフレームワークによって検証



検証FW：概要

- LLMのhallucination検知性能を測るフレームワーク
- 文中の知識が (subject, relation, object) の三組み (トリプル) で表される文から構成されるデータセットを使用
 - トリプルの例 : (Paul Mounsey, place of birth, Scotland)
- hallucination検知をある文の真偽判定タスクとして定式化し、生成文の真偽判定精度を測定
- 検証ステップ (詳細は後述)
 1. 真実の文の生成
 2. 虚偽の文の生成
 3. 生成文の真偽判定

検証FW : 1. 真実の文の生成

- データセット内の各文の言い換え文を生成
- objectフレーズを使うように指示
- GPT-3.5 Turboを用いた実験では、人手評価で99%が指示通りに生成されたことを確認

(Paul Mounsey, place of birth, **Scotland**)

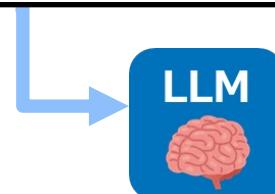
入力

Generate a paraphrase of the input. You must use the keyword:

Input: NahooToo is the second album by Scottish musician Paul Mounsey.

Keyword: **Scotland**

Paraphrase:



検証FW : 1. 真実の文の生成

- データセット内の各文の言い換え文を生成
- objectフレーズを使うように指示
- GPT-3.5 Turboを用いた実験では、人手評価で99%が指示通りに生成されたことを確認

(Paul Mounsey, place of birth, **Scotland**)

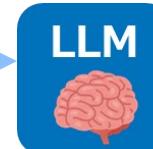
入力

Generate a paraphrase of the input. You must use the keyword:

Input: NahooToo is the second album by Scottish musician Paul Mounsey.

Keyword: **Scotland**

Paraphrase:



(Paul Mounsey, place of birth, **Scotland**)

出力

Paul Mounsey released NahooToo as his second album as a musician from **Scotland**

検証FW：2. 虚偽の文の生成

- ・ステップ1で生成した真実の文のobjectフレーズを書き換えて虚偽の文を生成
- ・一見すると本当に見える嘘の文を生成するように指示
- ・GPT-3.5 Turboを用いた実験では、人手評価で99%が指示通りに生成されたことを確認

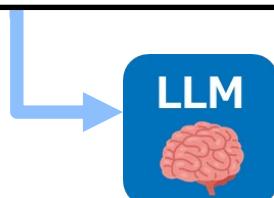
入力

Generate a false sentence that seems true. You must rewrite only the one keyword in the input:

Input: Paul Mounsey released NahooToo as his second album as a musician from Scotland.

Keyword: Scotland

False sentence:



(Paul Mounsey, place of birth, Scotland)

検証FW : 2. 虚偽の文の生成

- ・ ステップ1で生成した真実の文のobjectフレーズを書き換えて虚偽の文を生成
- ・ 一見すると本当に見える嘘の文を生成するように指示
- ・ GPT-3.5 Turboを用いた実験では、人手評価で99%が指示通りに生成されたことを確認

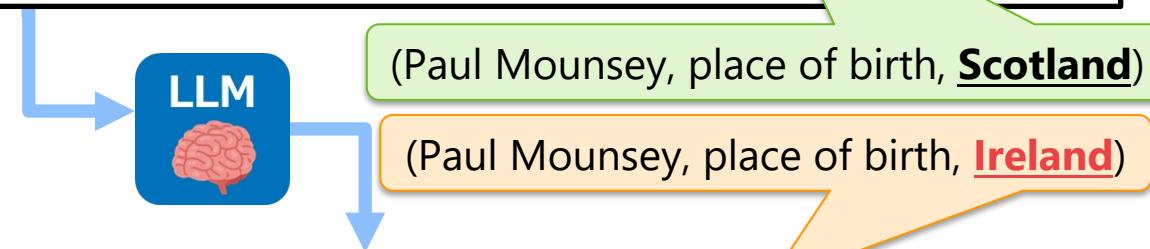
入力

Generate a false sentence that seems true. You must rewrite only the one keyword in the input:

Input: Paul Mounsey released NahooToo as his second album as a musician from Scotland.

Keyword: Scotland

False sentence:



出力

Paul Mounsey released NahooToo as his second album as a musician from Ireland.

検証FW : 3. 生成文の真偽判定

- ステップ1, 2で生成した文の真偽を判定
- 回答は "yes / no" で生成

入力 (真実の文)

Does the input contain misinformation? You must answer with yes or no:

Input: Paul Mounsey released NahooToo as his second album as a musician from **Scotland**.

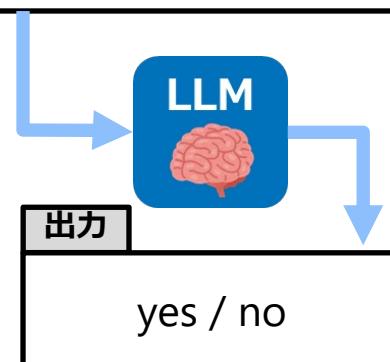
Answer:

入力 (虚偽の文)

Does the input contain misinformation? You must answer with yes or no:

Input: Paul Mounsey released NahooToo as his second album as a musician from **Ireland**.

Answer:



検証FW : CoTを用いた真偽判定

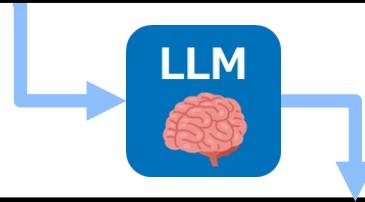
- LLMの知識はパラメータに埋め込まれており [4] , 真偽判定にはパラメータに内包されている知識の活用が重要
- パラメータから知識を上手く引き出すためにCoTを用いる

入力

Does the input contain misinformation? Let's think step by step to answer this question. You must answer with yes or no:

Input: Paul Mounsey released NahooToo as his second album as a musician from **Ireland**.

Thought:



文を意味のまとまりごとに分割し、根拠となる事実を生成して判定

出力

1. "Paul Mounsey released NahooToo as his second album": This part of the sentence is accurate. "NahooToo" is indeed the second album released by Paul Mounsey.

2. "as a musician from Ireland": **This part of the sentence contains misinformation. Paul Mounsey is not from Ireland; he is a Scottish musician and composer.**

So, the sentence contains misinformation in the second part where it incorrectly identifies Paul Mounsey as being from Ireland.

Answer: yes

実験：設定

- データセット：LAMAデータセット [5]
 - トリプルの情報が付与された英文で構成
 - NLTKで代名詞を含む文をフィルタリング
 - 41種類のrelationラベルがあり, 前処理後のテストデータは26,803文
- モデル：GPT-3.5 Turbo (2023/11/06 版)
 - 10-shotで生成
- 評価指標：recall, precision, F値, 正解率
 - hallucinationを見逃す偽陰性は偽陽性よりも深刻な誤りであるため, recallが最も重要な評価指標
 - recallはhallucination検知率とも呼ぶ

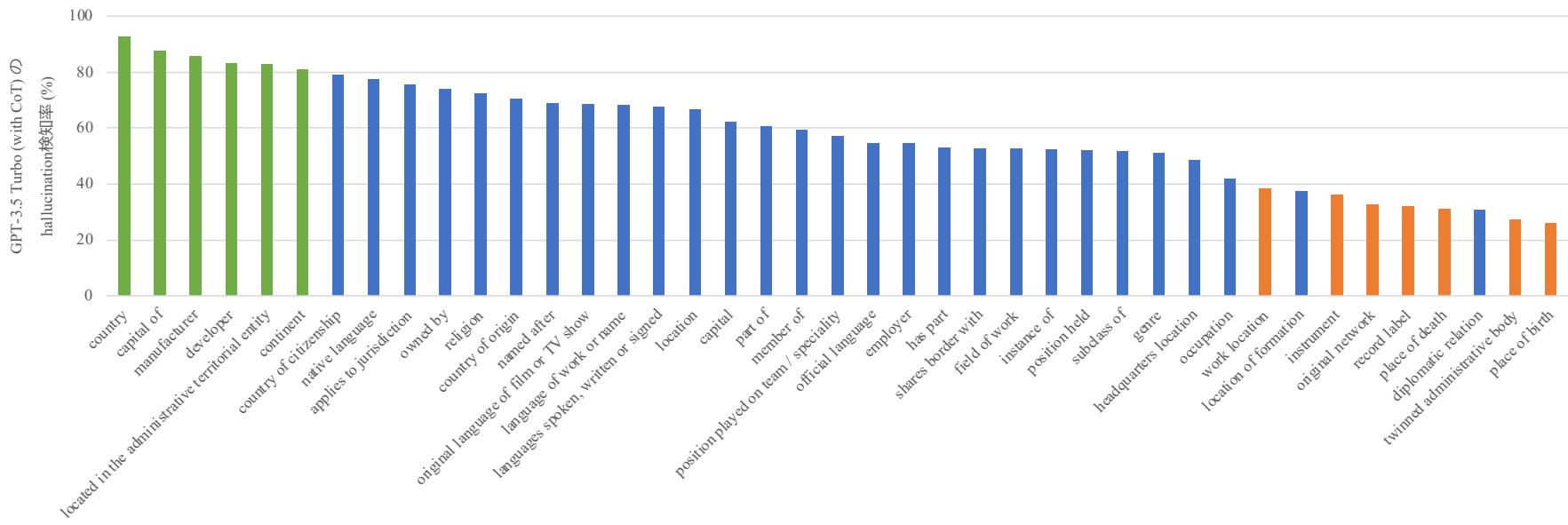
実験：結果

- CoTを用いない手法はrecallが21.9%しかなく, hallucination 検知器として使うことは難しい
- CoTによってパラメータに内包されている知識を引き出すことで, recallが36.3%, F値が33.8%, 正解率が14.5%向上
- precisionが若干低下した原因是, 事実や表現が曖昧な部分も誤情報として検知するようになったからだと考えられる

	recall (%)	precision (%)	F値 (%)	正解率 (%)
GPT-3.5 Turbo	21.9	85.1	34.9	58.0
GPT-3.5 Turbo (with CoT)	58.2	83.8	68.7	73.5

分析：relationラベルごとの検知率

- 得意 / 不得意 な分野が存在することが判明
 - 地理や企業に関する分野は80%以上検知できる
 - 人物やエンタメに関する分野は40%も検知できない
- LLMのパラメータに内包されている知識量が関係している?
→ hallucination検知率と知識量の関係を分析



分析：検知率と知識量の関係 (1/2)

- 背景

- LLMの知識量は事前学習コーパスに関連するテキストがどれだけ含まれているかに依存 [6]
- あるエンティティに関連するテキスト量と popularity には相関があると考えられる [7]
- hallucination検知率とpopularityの関係を分析

該当するWikipediaページの閲覧数と定義

- 設定

- テストデータをBINに振り分け、BIN内のhallucination検知率と平均popularityの Spearmanの順位相関係数 を求める
- 振り分け方
 - popularityに従って20個のBIN
 - relationラベルに従って41個のBIN

[6] <https://aclanthology.org/2022.findings-emnlp.59/>

[7] <https://aclanthology.org/2023.acl-long.546/>

分析：検知率と知識量の関係 (2/2)

- 結果
 - いずれの振り分け方でも正の相関あり
 - popularityに従って振り分けた際の相関係数は0.574で、有意に高い
- 結論
 - LLMのhallucination検知率は知識量に関係しており、パラメータに十分な知識が内包されていれば検知できる可能性が高い
 - relationラベルごとのhallucination検知の難易度はpopularityからある程度説明できる

振り分け方	相関係数	P値
popularity	0.574	0.008
relationラベル	0.337	0.031

まとめ

- RQ : **LLMは自身の*hallucination*を検知できるか？**
 - ある文の真偽判定タスクとして定式化し, LLMの*hallucination*検知性能を測るフレームワークによって検証
 - パラメータに十分な知識が内包されていれば, CoTにより知識を引き出すことで, 検知できる可能性が高い
- 本研究の貢献
 - LLMが自身の*hallucination*を検知できることを初めて定量的に示した
 - LLMの*hallucination*の改善に向けて, 各LLMがどのような分野で*hallucination*を起こしやすいかを評価する研究の足掛かり
 - *hallucination*検知とLLMの知識量に関連があることを発見し, 事前学習コーパスの量の重要性を改めて示した