FUJITSU

「AI（考える）」技術

AIのまちがいを見抜く技術 ～富士通のAIハルシネーション（幻覚）検出技術～

最終更新日 2025年3月28日

もくじ

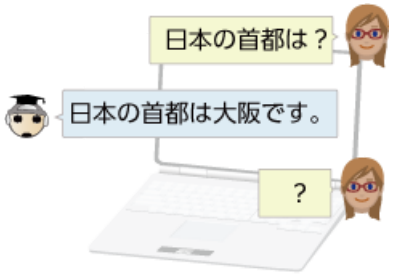
AIがまちがえる？！

AIはいつも正しいって思っていたんですけど？

AIも知らないことや勘違いすることもありますので、結果的にまちがった情報を回答してしまうことがあります。それを**AIハルシネーション（幻覚）**といいます。

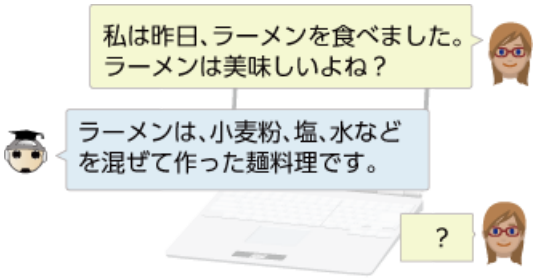
え！AIハルシネーション（幻覚）？

はい、生成AIが実際には存在しない情報を勝手に作り上げてしまう現象のことです。AIハルシネーション（幻覚）の例をだしてみましょう。
（生成AIが必ずしもこのような回答するわけではありません。生じ得る可能性がある極端な例のひとつです）



え？なんでこんな回答をしちゃうんだろう？

もしかしたらAIの学習データの中に「大阪は日本の経済の中心地」といった情報が多く含まれていて、「経済の中心地＝首都」と誤ってしまったのかもしれません。もうひとつ例をだしてみましょう。

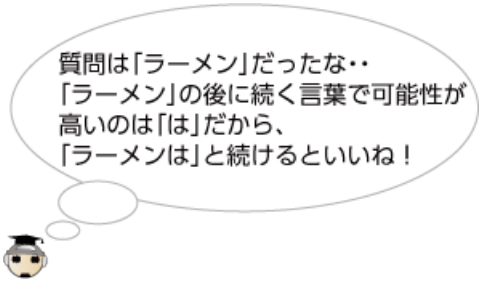


あれ？会話が成立していないですけど・・？

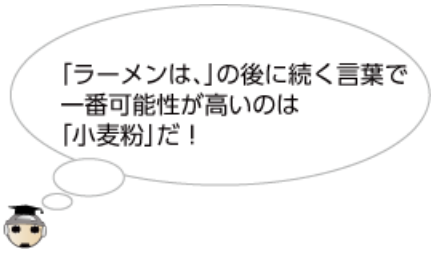
はい、ラーメンの説明としては良いかもしれませんが、質問への回答としてはまちがっていると言えます。つまり、これも**AIハルシネーション**です。

これもAIハルシネーションなんだ・・

生成AIは質問の意味を「理解」しません。
質問文の「ラーメン」という単語に注目します。「ラーメン」の次にくる可能性が一番高い単語を探します。
一番可能性が高い単語として**「ラーメン」＋「は」**になりました。



そして「ラーメンは」の次にくる可能性が高い単語を探します。
一番可能性が高い単語は**「ラーメンは、」＋「小麦粉」**になりました。



このように**生成AIは、次にくる単語を予測し続けることによって文章を作って**いるんです。

へ〜、知らなかった！

そのため、最終的にユーザーが質問した内容と、異なってしまうことがあります。このようなAIハルシネーションは、生成AIが文章を作る原理上、完全には防げません。

それは困りましたね。わかりやすい間違いならすぐに気がつけるけど・・・

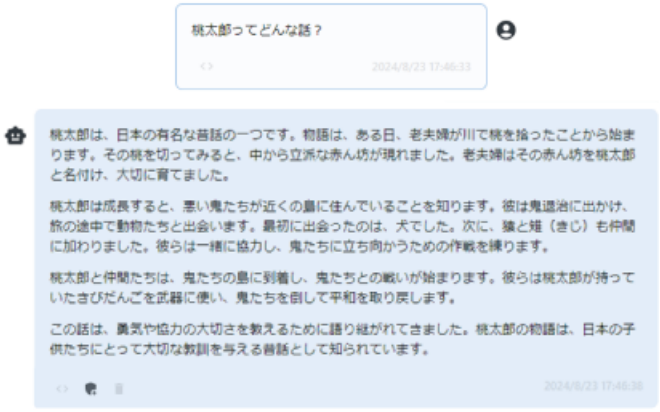
そうなんですよね。そこで、富士通では**まちがいをチェックするツール**を開発しました。実際のチェックツール画面を見ながら、この技術を説明しますね。

お願いします！

富士通のAIハルシネーション検出技術の実際の画面

昔話の「桃太郎」についてAIに質問して、その回答をチェックする画面を見てみましょう。

*この講座では、富士通の「Fujitsu Kozuchi 対話型生成AI」を使っています



チェックしたいAIの回答の左下に「盾」のマークがあります。そこをクリックすると、AIハルシネーション検出が始まります。

盾のマークをクリック

次にチェック方法を選択します。
今回は右側の「キーフレーズに注目して多数決」を選択し、右下の「Check」ボタンをクリックします。

回答全体を多数決
全体の幻覚スコアを表示します
結果を表示するのは速いですが
検出粒度が粗いです

キーフレーズに注目して多数決
文ごとに幻覚スコアがでます
遅いですが、検出粒度が細かいです

文章ごとにチェック結果として「幻覚スコア」が表示されます。**幻覚スコアとはAIハルシネーションである可能性の数値**で、「0」から「100」までの数値で表示されます。数値が大きいと、AIハルシネーションの可能性が高いことを意味しています。

「幻覚スコア」

数値が「0」なので、AI ハルシネーションの可能性が低いです

「幻覚スコア」

数値が「100」なので、AI ハルシネーションの可能性が高いです

今回のチェック結果の内、1つだけ「幻覚スコア」が「100」と表示された文があります。
「**彼らは桃太郎が持っていたきびだんごを武器に使い、鬼たちを倒して平和を取り戻します。**」という文です。

あれ？それは変ですね！

はい、この文章はAIハルシネーションの可能性が高い、という結果がでています。これを例文にして、従来技術の課題と富士通の新方法を紹介しますね。

お願いします！

世界トップクラスの技術を教えて！
～一般的な従来方法と富士通の新方法の違い～

・**従来方法：**正しい回答でもAIハルシネーションと判断されることがあり、正確さに欠けるのが課題

・**富士通の新方法：**従来方法よりも、正しい回答はきちんと「正しい」と判断する

AIが正しく回答したのに、その回答がAIハルシネーションと判断されるのは困りますね。

はい、そこが従来と新方法との「正確さ」の違いです。
それでは、例文を使って説明しましょう。

例文
「**彼らは桃太郎が持っていたきびだんごを武器に使い、鬼たちを倒して平和を取り戻します。**」

① 例文を文法的に解析し、「AIハルシネーション」が起こりやすい部分を自動的に空白にします

例文は
「彼らは（A）が持っていた（B）を武器に使い、 (C) を倒して平和を取り戻します。」となりました。

②　空白にした部分（A、B、C）に当てはまる言葉を複数回、自問自答します

自問自答って？

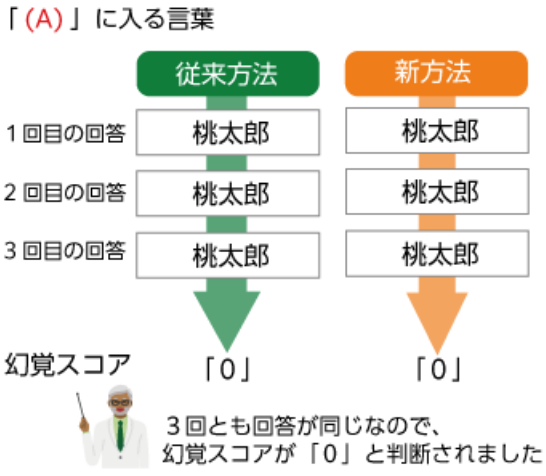
AIはこの例文に当てはまる言葉を、自分で自分に質問して回答を出してみます。

③ 当てはまる言葉が同じ回答なら「正しい」と判断とするため、幻覚スコアが「0」になります。異なる回答ならAIハルシネーションかもしれないので、幻覚スコアが高くなります。最大で「100」です。

A、B、Cのそれぞれの回答を順番にみてみましょう

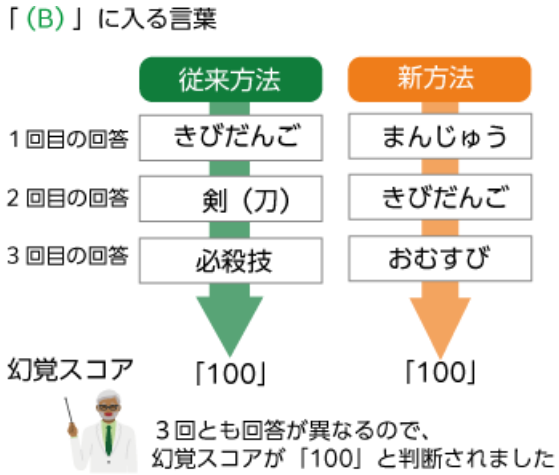
「彼らは桃太郎（A）が持っていた」

自問自答の結果：当てはまる言葉が3回とも「桃太郎」という回答だったので、根拠が得られたとして、従来方法も新方法も幻覚スコアが「０」と判断されました



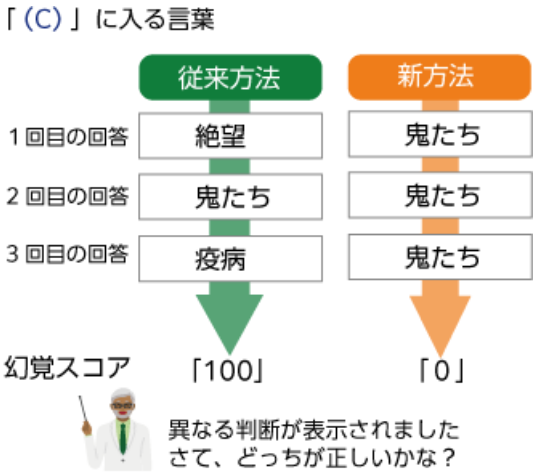
「きびだんご（B）を武器に使い」

自問自答の結果：当てはまる言葉が3回とも異なります。
十分な根拠が得られないと判断したため、従来方法と新方法とも、幻覚スコアが「100」になりました



「鬼たち（C）を倒して平和を取り戻します」

自問自答の結果：当てはまる言葉が従来方法の場合3回とも異なるので、幻覚スコア「100」になりました。新方法は、3回とも同じ答えなので幻覚スコアが「０」になりました。従来方法と新方法は、異なる判断をしました。



どうして異なる判断になったのでしょうか

従来方法は空白のところに入る可能性がある言葉がたくさんあるため、もともと答えがバラバラになりやすく、「AIハルシネーションかもしれない」と、**判断されやすいのです**。
新方法は空白を埋める**言葉の方向性をそろえるヒント**を自動的に追加しているので従来方法よりも正確に判断できます。

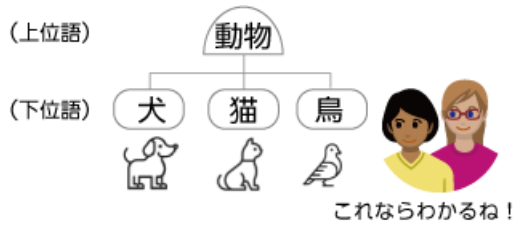
え、言葉の方向性をそろえるヒント？

はい、空白の部分に当てはまる言葉のヒント「**上位語**」と「**言語**（日本語や英語など）」を追加しています。

上位語？それはなんですか？急に難しくなった気がするのは私だけ？

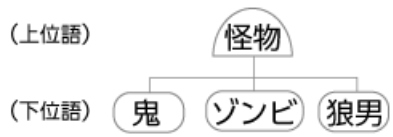
大丈夫ですよ、「上位語」とはある単語よりも広い意味を持つ単語のことです。例えば「犬」の上位語は「動物」です。

*上位語は、状況や目的によって変わります



「　(C)　」に入っていた「鬼」の上位語と言語は何になるのですか？

「鬼」の上位語はここでは「**怪物**」、言語は「**日本語**」になります。



上位語と言語が自動的にヒントとして追加されるため、従来方法の候補で、「**怪物**」ではない「絶望」や「疫病」は新方法では最初から候補外となります。

つまり、当てはまる言葉の範囲を適切に絞ることで、正しい答えの場合はきちんと「正しい」と判断できるってことですね。

その通りです。例えば家族で夏休みの計画をたてている時、自由に意見を言ったら希望がバラバラで決まらないということありますよね。ある程度方向性を決めてから希望を聞くとみんなの意見が一致しやすくなって決めやすい、ってことです。

【自由に意見を言った場合】

【ある程度方向性を決めて意見をいった場合】

なるほどね、この例えならわかりやすいわ！

この技術のすごいところは、空白のところを埋める言葉の方向性をそろえるための**ヒント（上位語＋言語）を自分（生成AI自身）で作って、自分に質問して自分で答えをだして判断している**ところです。

*生成した文章に否定文が含まれる場合も、同じ方法で判定することできます。
*生成した文章に含まれる挨拶文は省くように設定しています。

自分がだした生成文の内容が合ってるかどうか、それさえも自分で判断するなんて賢いね！

ちょっとまって！ファクトチェックと何が違うの？

「ファクトチェック」って、その記事が事実かウソかのチェックですよね。

「事実を検証する」のがファクトチェックです。

富士通のAIハルシネーション検出技術も幻覚か否かをチェックするのだから、その違いがわからないわ。

その2つは使い方が違います。ファクトチェックは、「web上にすでに投稿された記事の内容がウソかホントか」のチェックです。AIハルシネーション検出技術は、個人が使う生成AIからの回答をその場で幻覚か否かを判断する技術で、AIの回答をどう使うかは、ユーザーが決めます。

つまりファクトチェックはすでにネット上に誰かが公開した情報をチェックするもので、AIハルシネーション検出技術は自分が生成AIから得た情報のチェックってこと？

はい。ざっくり、そのように思ってもらえれば良いと思います

どんな人が使うと便利なの？

色々な職業の人に使ってもらえる技術です。紹介しますね。

・編集者

AIが生成した文章の誤りを見つけ出し、正確で信頼性の高いコンテンツを出版できます。

・カスタマーサポート

お客様の障害対応としてAIチャットボットが生成した回答の誤りを検出し、より正確な情報提供できるようにチェックするのに使うと、顧客満足度の向上につながります。

・サービスや商品に関するウェブサイトなどのFAQを作る人

AIが生成したFAQの内容を分析し、誤った情報が含まれていないかを確認するのに役立ちます。

AIハルシネーション検出技術はあくまでもツールであり、最終的な判断は人間が行う必要があることをおぼえておいて下さいね！

はい!

小話

・AI関連の技術開発はスピードが大切！

今回のハルシネーション検出技術は、生成AIで使っているLLM（大規模言語モデルでテキストを生成する技術）の特性に気づいてからわずか4か月で社外発表をすることができました。

ええー！開発期間ってそんな短かったんですか？

はい、開発にとても時間を要するものもありますが、AIについては進化がとても速いので、アイディアを速く形にしないと、他の研究者に先を越されてしまうんです。
実はこんなに短期間でつくれたのは社員の「フィードバック」のおかげです。
富士通の社内掲示板に新技術を試してほしい、と記事を投稿すると新しいものが好きな社員がすぐに試してくれます。そしてうまくいかない例をフィードバックしてもらって。どんどん解決しました。

それで、たった4か月で社外発表できたんですね！

・赤ちゃんとともに♪

研究員のMさん、実はまだゼロ歳児のパパなんですよね。

はい、会社の子育て支援の「サポート休暇」を取得して、復職してからすぐにアイディアを思いついて開発を始めました。ただ、ゼロ歳児ということもあり、自分が子供をみる時間帯が必要でした。

それは大変ですね！仕事と子育てをどのように両立させているんですか？

今は、自宅テレワークで近くに子どもがいる環境で仕事をしています。そして打合せが入った時は、事前に状況を周知しています。打合せ中に子どもがグズリだすと、あえてカメラをONにしたりして、場が和やかになったりするんですよ。

そういう雰囲気って大事ですよね。それに、お互いの状況を受け止め合えることで、さらにポジティブに仕事を進められますよね！

はい♪

* 写真はイメージです。研究員Mさんではありません

関連リンク

プレスリリース

- 対話型生成AIの幻覚やAIを騙す敵対的攻撃に対処できるAIトラスト技術を開発し、「Fujitsu Kozuchi (code name) - Fujitsu AI Platform」で提供開始

論文（arXiv:2409.17173)

- A Multiple-Fill-in-the-Blank Exam Approach for Enhancing Zero-Resource Hallucination Detection in Large Language Models

IP Innovation Award 2024

- 富士通の知財活動の促進に大きな貢献がある社内活動を表彰
  IP Innovation Award 2024 | Public Detail | Open Badges Wallet

Copyright 1994 - 2025 Fujitsu