LLMによる要約・詳細化過程におけるハルシネーションの分析

山田 夏稀[1]　安尾 萌[2]*　松下 光範[3]　Junjie Shan[1]　西原 陽子[1]
Natsuki Yamada[1]　Megumi Yasuo[2]　Mitsunori Matsushita[3]　Junjie Shan[1]　Yoko Nishihara[1]

[1] 立命館大学情報理工学部
[1] College of Information Science and Engineering, Ritsumeikan University
[2] 立命館グローバル・イノベーション研究機構
[2] Ritsumeikan Global Innovation Research Organization
[3] 関西大学総合情報学部
[3] Faculty of Informatics, Kansai University

Abstract: 本研究の目的は，大規模言語モデル（以下，LLM）が要約と詳細化を繰り返す過程で情報がどのように変化し，ハルシネーションが生じるかを明らかにすることである．ニュース記事を用い，LLMによる要約・詳細化を複数回繰り返し，出力結果をハルシネーションのタイプ別に分類，原文および出力結果の修辞構造の観点から分析した．その結果，具体的な数値や根拠情報が失われるなど，既存研究で報告されている人のうわさの伝搬行為と同様の変化が確認された．また修辞構造の観点では，LLMがハルシネーションを含めた上で，元となる記事と類似した文章構成のテキストを生成することが確認された．

1　はじめに

Web上の偽情報や誤情報の蔓延は，社会的な分断や意思決定の誤りを引き起こす深刻な問題となっている．誰もが簡単に情報発信が可能な現代において，その拡散速度と影響範囲は増大し続けている．こうした偽情報や誤情報は，発信者個人の悪意や誤解のよって生じるものに限るものではなく，情報の伝搬プロセスの過程でも生じえる．情報が伝播するプロセスの一例として伝言ゲームがある．伝言ゲームは情報が人から人に経由していくことで，最初の情報が変化・劣化する様子を楽しむゲームである．伝言ゲームのプロセスは，もととなる情報の伝達と，伝達の際に抜け落ちた情報の補完の繰り返しとして捉えることができ，実際のWeb上での情報の拡散プロセスの例として用いられることがしばしばある．

LLMの登場により，情報の生成や要約が自動かつ高速に実行可能となった．実際に，LLMを用いてWeb上のニュース記事を要約し，報道内容の概略を投稿するボットなどがソーシャルメディア上で運用されている．しかし，LLMもまた事実と異なる情報を生成する現象（以下，ハルシネーション）を引き起こし，誤った情報を拡散させる懸念がある．Webクローラによる高速な情報収集と，LLMによる高速な情報の生成および要約が組み合わさることで，LLMを介した伝言ゲームは，従来とは比較にならない速度と規模で誤情報を生成・拡散させるという新たな脅威となりうる．

LLMの誤情報拡散は，もととなったニュース記事からLLMが内容を要約し，要約された内容から元の内容が補完される，という処理の繰り返しの過程でハルシネーションが混入し，それが拡散されることで発生すると考えられる．本研究では，LLMを介した伝言ゲームのプロセスを「要約」と「詳細化」の2つの処理の繰り返しと捉える．従来の誤情報研究は，人間の認知バイアスや社会的拡散パターンを中心としてきた[1]．一方で，LLMによって生成される誤情報の段階的な変化に焦点を当てた研究は少なく，要約と詳細化のサイクルを繰り返す中で，元の情報が具体的にどのようなメカニズムで変化し，どの段階でハルシネーションへと劣化していくのか，そのプロセスは解明されていない．本稿では，LLMに要約と詳細化からなる伝言ゲームのプロセスを実行させ，その過程で観察されるハルシネーションの特徴を分析する．

2　関連研究

2.1　LLMとハルシネーションに関する先行研究

Huangらは，ハルシネーションに関する包括的なサーベイを提供し，ハルシネーションが要約タスクにおいて

*連絡先：立命館大学情報理工学部
〒 567-8570 大阪府茨木市岩倉町 2-150
E-mail:{yasuo-ri,nisihara}@fc.ritsumei.ac.jp

も深刻な影響を及ぼすことを報告している [2]．Maynez らは，大規模な人手評価を行い，当時の最先端の要約モデルであっても，生成された要約の多くが入力記事に忠実でない情報を含んでいることを明らかにした [3]．Kalai らはハルシネーションの発生原因について，モデルが不確実な場合に推測で回答しても，人手評価や自動指標によって報酬が与えられる設計になっていることが，その一因であると指摘している [4] ．

2.2 人間社会における伝言ゲームでの誤情報伝搬

情報の伝播が人々の行動に与える影響は，古くから社会の重要な関心事であった．Allport らは噂の伝達実験を行い，伝達する過程で内容が短く平易になる「平均化」，特定の要素が強調される「強調化」，人の既存の知識や信念に沿うように内容が歪められる「同化」，情報に要素や詳細が追加される「付加化」といった現象が起きることを実証した [5]．このような人間の認知的バイアスによる情報変化の知見は，災害時などの現実社会での情報伝播の分析にも応用されている．小笠原らは，災害時の情報伝播に関する研究において，人々が曖昧な状況に対して独自の解釈を追加することで不安を反映した誤情報の種が生まれるプロセスを示した [6]．誤情報が社会パニックを引き起こした実証事例として，有馬らは 1973 年に発生した豊川信用金庫の取り付け騒ぎを詳細に分析している．この事例は，誤情報が社会パニックへと発展するプロセスを記録している．この調査では，豊川信用金庫に就職が決まった友人に対して発された「信用金庫は危ない」という発言が伝播する過程で，主語が曖昧な「信用金庫」から「豊川信用金庫」へと対象が特定化され， さらに，本来の意図とは異なる「経営が危ないらしい」という推量の噂へと意味内容が変化し，最終的には潰れるという根拠のない断定へとエスカレーションしたというプロセスが示されている [7]．これらの先行研究から，情報は伝達過程において，不確実な状況に関する解釈が付け加えられることで誤情報が増加し，社会的なパニックを引き起こしうることが示されている．

2.3 本研究の位置づけ

LLM による情報の反復的な処理は，人間が情報を伝達する伝言ゲームのプロセスと構造的な類似性を持つ．たとえば，ある事象についての内容を簡潔に伝える処理，および要約によって省略された内容を復元する処理を繰り返すことは，あるニュースのタイトルや趣旨のみを伝達し，その内容について事後的に肉付けされるという構造と類似する．LLM の情報処理プロセスが社会に与える影響は増大しつつあることから，豊川信用金庫の事例 [7] や災害時における誤情報伝播で観察される同化・付加化 [6] といった，情報伝達の変化が，LLM においても再現されるのか，あるいは LLM 固有の全く異なるエラーパターンが出現するのかを明らかにすることは，LLM が社会に及ぼす影響を予測する上で不可欠な研究課題である．

以上を踏まえ本研究では，LLM の反復的な情報処理を，人間の伝言ゲームにおける情報の変化と対応させて分析する．内容を要約するプロセスと，元の要約を復元するプロセスを反復するシミュレーションを行い，単一のエラーを特定するだけでなく，サイクルを通じてハルシネーションが蓄積，変化するプロセスを分析することで，LLM が情報の再生成を繰り返す中で，どのような変化パターンを示すかを定性的に解明する．

3 分析方法

3.1 LLM を用いたハルシネーションを含むデータの生成

本研究では LLM が生成する可能性があるハルシネーションを分析するため，人間社会の噂の伝播モデル [5] から，伝達する過程で内容を短縮するプロセスと，短縮された内容を再度復元するプロセスに着目し，LLM が生成するハルシネーションを，この 2 つのプロセスからなると仮定した．本稿では各プロセスをそれぞれ「要約」および「詳細化」と記述する．この 2 つのプロセスを LLM に反復的に実行させることで，ハルシネーションを含む情報を生成させる．

LLM を用いてハルシネーションを含むデータを生成させる手順について説明する．

1. **要約ステップ**：原文，または 1 つ前の詳細化で得られたテキストに対し，LLM を用いて要約をする．要約の文字数は元のテキストの 50%程度と指定した．
2. **詳細化ステップ**：要約ステップで得られたテキストに対し，同じく LLM を用いて詳細化を行う．詳細化の文字数は原文テキストと同程度と指定した．

この調査では，上記の要約ステップと詳細化ステップを各 10 回繰り返した．結果として，原文テキスト 1 件に対し，要約テキストが 10 件，詳細化テキストが 10 件得られた．本論文で使用した LLM は GPT-4（gpt-4-0613）であった．

原文テキストは，Yahoo!ニュースに掲載されたニュース記事を対象とした．Yahoo!ニュースはニュース記事のカテゴリを国内，国際，経済，エンタメ，スポーツ，

IT，科学，ライフ，地域に分類している．この実験ではYahoo!ニュースにおけるニュース記事のカテゴリ分類に基づき，9つのカテゴリからそれぞれニュース記事を1件ずつ取得し，分析に用いた．記事の取得日は2025年4月8日，4月12日，および5月29日であった．

3.2 ハルシネーションの種類分析

生成された詳細化テキストに対し，ハルシネーションの種類を分析する．本研究でのハルシネーションの定義は,「原文テキストに存在しない，あるいは文脈と矛盾する内容が生成される現象」とする．ハルシネーションの内容は既存研究を参考にし，以下の5種類とする [2].

1. **外在的事実誤り**：訓練データに含まれない事実誤り.（ex. 原文にない誤った年号や固有名詞の生成）
2. **内在的事実誤り**：学習に用いたデータには存在する情報だが，現実とは異なる事実を出力する誤り.
3. **指示不一致**：「詳細化する」という指示に対し，原文にはなかった独自の解釈や評価（ex. これは世界経済の健全さを示している）を生成する誤り.
4. **文脈矛盾**：会話履歴や前段の生成内容と矛盾する誤り.
5. **論理的不整合**：生成された詳細化文の内部で，論理的な矛盾（ex. 前半と後半で主張が異なるなど）を引き起こす誤り.

内容分析では，詳細化テキストを1文ずつ人手により評価する．ある文が以下の3つの場合分けのいずれかに該当する場合は，ハルシネーションが含まれる文が生成されたと評価する.

1. 原文テキストに含まれていない情報を含む.
2. 原文テキストの内容と矛盾する記述を含む.
3. 原文テキストの文脈からは支持されない事実に基づかない推論や断定を含む.

ハルシネーションを含む文に対し，先に示した5種類のラベルのいずれかを付与する．ラベルは複数の付与を可能とした．LLMがもととなる記事からハルシネーションを生成する際の変化を観察するため，10回実施された要約–詳細化ステップのうち，初期段階である1回目と2回目のステップで生成された詳細化テキストを対象として内容分析を行う.

3.3 修辞構造タグを用いたハルシネーションの構造分析

LLMにより要約と詳細化が繰り返されることで生成されるテキストについて，構造レベルでのハルシネーションを分析するために，修辞構造タグを用いて分析を行う．修辞構造タグを用いることでテキストの構造が把握できるため，LLMが要約と詳細化を行う際に原文テキストの構造を認識しているのか，また構造を保持した上で生成を行っているのかを確認する．構造変化が原文の持つ固有の特性に強く依存するのであれば,類似したジャンル同士はその変化のパターンもまた類似するという仮説を立てた.

分析では，詳細化された1回目から10回目までの10種類のテキストに対し，テキストに含まれる各文に対し，修辞構造を示すタグを付与した．修辞構造を示すタグは既存研究に基づき [8],「原因」,「条件」,「否定条件」,「目的」,「譲歩」,「対比」,「例外」,「類似」,「代替」,「連言」,「選言」,「例示」,「詳細化」,「言い換え」,「同時性」,「非同時性」,「展開」,「評価」の18種類のタグを使用した.

タグの付与手順はLLMを用いて以下の手順で行った．まず，タグの定義と例をプロンプトとして用意し，詳細化テキストの文と合わせてLLMに入力する．LLMが付与したタグについて，第一著者が原文の文脈と照らし合わせ，修正を行った上でタグ付与を完成させた．使用したLLMはGPT-4（gpt-4-0613）であった.

続いて，付与されたタグを詳細化のテキストごとにベクトル化する．ベクトルの各要素を修辞構造を表すタグとし，テキストに含まれる修辞構造のタグの割合をベクトルの値とする．1つの詳細化テキストに対し,18次元のベクトルが得られる．1つのジャンルごとに18次元のベクトルを，1回目から10回目までの詳細化テキストに付与されたタグの割合をステップ順に連結し，$18 \times 10 = 180$ 次元のベクトルを得る.

続いて，ベクトル間の類似度を算出する．異なるジャンルの2つのベクトルの類似度をコサイン類似度により算出する．これにより，9ジャンル，$9 \times 8/2 = 36$ 個のコサイン類似度が得られる.

4 分析結果

4.1 ハルシネーション内容分析の結果

ハルシネーションの内容分析の結果を表1に示す．内容分析の結果,「指示不一致」,「外在的事実誤り」,「内在的事実誤り」,「文脈矛盾」,「指示不一致+外在的事実誤り」,「指示不一致+内在的事実誤り」,「指示不一致+文脈矛盾」，の7種類のタグとその組合せが観察された．最も多く観察されたハルシネーションは「指示不

表 1: ハルシネーション分類の集計

行ラベル	IT	エンタメ	スポ	ライフ	化学	経済	国際	国内	地域	総計
指示不一致	3	5	2	4	9	3	2	5	8	**41**
外在的事実誤り	0	1	0	0	0	0	2	1	0	4
内在的事実誤り	1	0	2	4	1	1	7	1	0	**17**
文脈矛盾	0	0	0	0	0	0	0	0	1	1
指示不一致＋外在的事実誤り	0	2	0	0	0	0	0	0	0	2
指示不一致＋内在的事実誤り	0	0	0	0	0	1	1	0	0	2
指示不一致＋文脈矛盾	0	0	1	0	0	0	0	0	0	1
総計	4	8	5	8	10	5	12	7	9	68

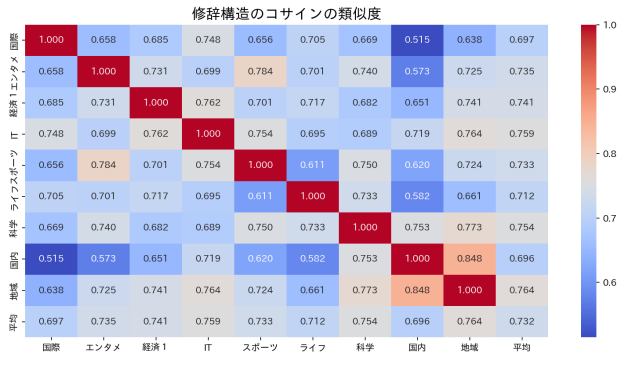

図 1: 修辞構造タグの推移を表すベクトルの類似度を示したヒートマップ．類似度はコサイン類似度．ジャンルごとにベクトルを作成し，ジャンル間の類似度を算出した．

一致」であった．続いて「内在的事実誤り」のハルシネーションが多くなった．一方,「論理的不整合」に該当するハルシネーションは観察されなかった．この原因として，今回の調査において入力されたニュース記事が LLM のコンテキストウィンドウの範囲内であり，一貫性のある文章が生成可能な状況であったこと，また LLM が文章生成において文法や意味的なつながりを優先するため，文内部で矛盾を回避する傾向にあるためだと推測できる．

4.2 修辞構造タグを用いた分析の結果

修辞構造タグを用いた分析の結果として，ジャンル間のコサイン類似度をヒートマップとして図 1 に示す．類似度が高かったジャンルの組合せは，国内と地域（0.848），科学と地域（0.773），IT と地域（0.764）であった．類似度が低かった組合せは，国際と国内（0.515），エンタメと国内（0.573），ライフと国内（0.582）であった．また，IT，地域，エンタメ，科学は他ジャンルとの平均的な類似度が高かった．

表 2: 詳細化テキストに含まれたハルシネーションの例．指示不一致の内，原文テキストにない情報の付加の例．

原文テキスト	詳細化テキスト
ラッパーの Awich がサプライズで登場し、和と融合したラップで盛り上げたり、歌舞伎役者・尾上菊之助らが舞を披露したりするなど、日本色溢れるステージで華を添えた。	また、コブクロや絢香、imase、尾上菊之助らも登場し、歌や演技で万博の開幕を祝いました。
シンガーソングライター・imase とともに、最新のテクノロジーと音楽が融合したステージで盛り上げ、会場の心をひとつにした。	これらの出演者は、日本の伝統文化と最新技術を融合させたパフォーマンスを披露し、**万博のテーマ「共生・創造・未来」を象徴する華やかな内容となりました。**
吉村洋文大阪府知事、同博覧会名誉会長の石破茂首相らがあいさつ。	式典では、石破茂首相や吉村洋文大阪府知事も挨拶を行い、**日本の進歩的な技術や文化を世界に向けて発信する決意を表明しました。**

5 考察

5.1 ハルシネーションの内容分析に関する考察

表 1 において，最も観察された回数が多かったのは「指示不一致」のハルシネーションであった．指示不一致の中でも頻繁に見られたのは，原文テキストにない情報の追加であった．この例を表 2 に示す．表 2 では，原文テキストは万国博覧会の開会式に関する記事であった．1 回目の詳細化において，原文テキストには含まれないが，万国博覧会のテーマに関する記述が追加された（表 2 中太字箇所）．追加された原因としては，LLM が詳細化を実行する際に原文テキストの情報を忠実に保持することにより，より流暢で尤もらしいテキストの生成を優先した結果と考えられる．また，このハルシネーションは，人間による噂伝播モデルで指摘される同化および付加化のプロセスとも類似している．具

体的には，原文テキストのトピックから連想される情報の追加（万国博覧会のテーマ追加）は，読み手の既存知識に沿うように内容が歪められる同化と類似していると考えることができる．

さらに，原文テキストに含まれている重要な情報，特定の数値データや統計情報などが，詳細化のテキストでは失われることが多かった．他には，原文テキストで記述されていた事実や発話の内容を変更する現象も確認された．スポーツの監督や選手によるインタビュー記事において，その発言のニュアンスや内容が，原文テキストの内容を忠実に復元せず，異なる形に置き換えられていた．

5.2 修辞構造タグを用いた分析結果に関する考察

ジャンル間の修辞構造タグの類似度について，類似度のスコアが高かった組合せは，国内と地域，経済とエンタメ，IT と地域であった．これらの組合せに対し，原文テキストを段落単位に分割し，全体の構成を定性的に比較したところ，原文テキストの文章構成と論理展開がジャンル間で類似する傾向が見られた．表 3 は，国内ジャンルの記事で扱われた関西万博開幕時の喫煙問題と地域ジャンルの記事で扱われた愛知県の特殊詐欺リクルーター摘発事件について，両者の段落構成を比較した例である．両記事は扱うテーマこそ異なるものの，時系列に沿った説明，中盤での前提情報の挿入，関係者のコメント提示，そして最後に背景や動機を述べるといった構造が共通していることがわかる．一方で，要約と詳細化を繰り返す中で，特定の修辞構造の単調増加や単調減少の傾向は見られなかった．

図 2 に，原文と 1 回目から 10 回目までの詳細化で見られた修辞構造タグの出現回数の推移を示したグラフを示す．2 つの図からは，修辞構造タグが要約と詳細化が 1 回行われるごとに，含まれる修辞構造タグの数に変化があることが示されている．これは，LLM が生成するハルシネーションを自然な文脈のテキストに組み込むために，新たな修辞関係を生成した結果と考えられる．実際に，表 4 に原文と詳細化の文を比較し，ほぼ同一の内容で修辞構造が変化した例を示す. この例では，関西万博の開会式において登場したゲストとパフォーマンスについて説明するという同一の内容を記述しているものの，修辞構造が変化していることが確認できる．

ジャンル間で類似度の高い組み合わせがあるという結果は，ジャンル間での修辞構造の変化の仕方が類似していたということを示している．修辞構造の変化の仕方が類似した原因として，原文テキストの文章構成と論理展開が似ていた可能性が考えられる．したがって，LLM による要約と詳細化の繰り返しによって生成された文章は，原文テキストの文章構成に依存して構造が類似することが示唆される．

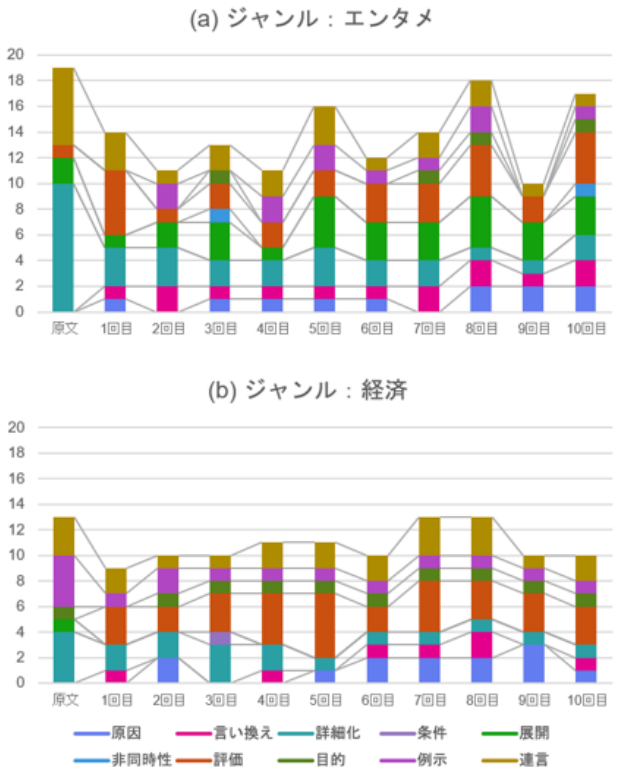

図 2: 修辞構造タグの出現回数の推移

6 おわりに

本研究では，LLM による要約と詳細化が繰り返されることで生成されるハルシネーションの分析を行なった．分析のために，ニュース記事を原文テキストとして，LLM を用いて要約と詳細化を繰り返してハルシネーションが含まれるテキストを生成した．生成されたテキストのうち，詳細化の段階で得られたテキストと原文テキストを比較することにより，ハルシネーションの種類を分析した．

分析の結果，生成されるハルシネーションとしては，(1) 原文テキストにはない情報が追加される，(2) 数値データ，統計データが欠落される，(3) 発言のニュアンスや内容が書き換えられる，などが多いことがわかった．さらに，各テキストに含まれる文に対し，修辞構造の分析を行った結果，ハルシネーションを含めた上で自然な文脈のテキストを生成するために，原文テキストの文章構成を踏まえて，類似した文章構成のテキストを生成することが分かった．

表 3: 国内ジャンルと地域ジャンルにおける段落構成の比較（類似度 = 0.848）

国内	地域	共通する構成特徴
時系列進行，途中で前提（ルール）を挿入，末尾に背景を提示	時系列進行，途中で兆候や動機を補足，末尾に背景・動機を提示	時系列ベース，中盤に前提／兆候，末尾に背景（動機）を配置
第 1 段落：万博開幕と状況説明	第 1 段落：摘発の事実を提示	いずれも事実提示で開始
第 2–3 段落：喫煙行為と「全面禁煙」という前提提示	第 2–3 段落：不審行動と，それを察知できた理由（兆候）	中盤で前提や兆候を示す構成
第 4–6 段落：違反状況，違反者コメント，運営側対応	第 4–8 段落：母親の行動，警察対応，供述，事件詳細	中盤〜後半で行動・供述の詳細を提示
第 7 段落：過去事故への言及（背景）	第 9 段落：母親の動機と警察の呼びかけ	最終段落に背景・動機を配置

表 4: 詳細化による修辞構造の変化例（情報の分割と評価の付与）

原文テキスト	詳細化テキスト
ラッパーの Awich がサプライズで登場し、和と融合したラップで盛り上げたり、歌舞伎役者・尾上菊之助らが舞を披露したりするなど、日本色溢れるステージで華を添えた。 **タグ：**詳細化， 評価	サプライズゲストとして、ラッパーの Awich が登場し、そのパワフルなパフォーマンスが会場を盛り上げました。 **さらに**、歌舞伎役者の尾上菊之助らが**美しい**舞を披露し、伝統的な日本の芸能が見られるなど、日本色溢れるステージが繰り広げられました。 **タグ：**連言， 評価

今後は，LLM ごと，あるいはニュース記事ジャンルごとに追加の分析を行い，結果を比較することで，発生するハルシネーションの種類とその伝達プロセスの違いについて明らかにする．また，修辞構造の順序や構造について追加の分析を行い，構造レベルでの情報の変化を明らかにする．

参考文献

[1] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[2] 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, 2023.

[3] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919, 2020.

[4] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025.

[5] Gordon W. Allport and Leo Postman. *The Psychology of Rumor*. Henry Holt and Company, 1947.

[6] 小笠原 盛浩, 川島 浩誉, and 藤代 裕之. マスメディア報道は Twitter 上の災害時流言を抑制できたか？—2011 年東日本大震災におけるコスモ石油流言の定性的分析. **関西大学社会学部紀要**, 49(2):121–140, 2018.

[7] 有馬 守康, 齋藤 哲哉, 小林 創, and 稲葉 大. 「取り付け騒ぎ」に関する理論的・実験的分析と事例との整合性に関する考察. **日本大学経済学部経済科学研究所紀要**, 49:45–53, 2019.

[8] Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 549–558, 2020.