大規模言語モデルによる言語生成における幻覚について

Hallucinations of the generation by Large Language Models

元山梨英和大学　教授
江原　暉将

1967 年早稲田大学理工学部卒。同年 NHK 入局。2003 年 諏訪東京理科大学教授。2009 年 山梨英和大学教授。2015 年 退職。アジア太平洋機械翻訳協会 (AAMT) ／ Japio 特許翻訳研究会オブザーバ。

eharate@gmail.com

① はじめに

2022 年 11 月に ChatGPT が公開されて以来、大規模言語モデル (Large Language Model: LLM) を用いた対話システムが急速に進歩している。2 言語以上の言語データで訓練された LLM は機械翻訳能力も有しており精度も高い。

LLM による言語生成は自然性が高く、人との対話であるかシステムとの対話であるかを区別するのが困難なレベルになっている。しかしながら高い自然性がそのまま高い正確性を意味するわけではなく、時折誤った内容を話すことがある。また誤りとは言えなくとも疑問に思う発言もある。このような発言は幻覚 (ハルシネーション) と呼ばれ、LLM を用いた言語生成における問題点の一つとなっている。

② LLM による言語生成における幻覚とは

幻覚の例を筆者の経験から紹介しよう。神社の本殿の両脇に比較的小さな社殿がある場合がある。右脇殿と左脇殿と呼ばれる。この左右の意味は、参拝者から見ての左右なのだろうか、それとも神殿に祭られている神様から見ての左右なのだろうか。Web 検索しても分からないので ChatGPT(無料版) に問い合わせてみた。その時[1]の対話の概要を以下に示そう。

(対話始)

江原　右脇殿とは本殿に向かって右側の社殿ですか。それとも本殿に向かって左側の社殿ですか。

GPT　「右脇殿」という用語は、場合によって異なる意味を持ち得ますが、一般的には本殿に向かって右側の社殿を指します。つまり、本殿の右側に位置する社殿のことを指します。

江原　右脇殿とは本殿に向かって左側の社殿ではないですか。

GPT　申し訳ありません、前回の回答が誤りでした。右脇殿は本殿に向かって左側の社殿のことを指します。つまり、本殿の左側に位置する社殿のことを指します。誤った回答をしてしまい、ご迷惑をおかけして申し訳ありませんでした。

(対話終)

このように前言を翻しており、最初の回答は幻覚といえる[2]。

LLM による言語生成における幻覚について文献 [1] では以下のように分類している。

・事実の不正確さ：例えば数値を間違える。正しい答は「患者の血糖値は 120 mg/dL である」に対して、LLM が「患者の血糖値は 150 mg/dL である」と誤って示す。

1　2023 年 1 月。

2　「場合によって異なる意味を持つ」と発言しているので、厳密には幻覚ではないのかもしれない。なお、2025 年 6 月に神社本庁の下部組織である東京都神社庁に問い合わせたところ「神社によって異なるが多くは神様から見ての左右である」とのことであった。

・**誤解**：LLM を訓練するコーパスや利用者の質問（プロンプト）を誤って解釈する。後者の例として”What is the meaning of lead?”という質問に対して”lead”の意味を「リーダシップ」と「鉛」の間で誤解する。

・**膨大な情報の中から必要な情報を見つけられない（干し草の山の中の針）**：例えば第一次世界大戦の原因を 1 つだけ挙げて、他の原因を省略する。

・**捏造**：架空の科学的研究を作成したり、実在しない歴史上の人物の引用をでっち上げたりする。

そして、文献 [1] では幻覚を引き起こす原因を訓練段階と利用段階（推論段階）に分けて分析している。利用段階の原因の一つとして推論の誤りがある。LLM による言語生成では、質問に対する回答を単に訓練データに含まれる言語表現から検索して出力するのではなく、推論を利用して生成している。この推論段階で幻覚が生ずる可能性がある。論理的推論には、演繹推論、帰納推論、逆行推論などがあるが、演繹推論における幻覚について次節で考察しよう。

③ 演繹推論における幻覚

演繹推論とは前提から出発して推論規則を適用して結論に至る道筋を得るものである [2]。記号論理学（古典命題論理）では【表 1】に示すような同一律、矛盾律（爆発律）、排中律を前提そのものあるいは推論可能な命題として用いている [2][3]。表中 A や B は任意の命題であり、→は「ならば」、∧は「かつ」、∨は「または」を表す論理記号である。また、¬ A は A の否定を表す。

古典命題論理は「真」と「偽」の 2 つの真理値をとる 2 値論理によって無矛盾性と完全性が証明されている [2]。

表 1　古典命題論理における 3 原理

同一律	$A \rightarrow A$
矛盾律	$(A \wedge \neg A) \rightarrow B$
排中律	$A \vee \neg A$

ここで LLM における推論を考察すると、推論の前提には訓練に用いた言語データベースに含まれる言明（命題）がある。それらの前提から推論可能な命題を推論結果として生成する。LLM における推論には表 1 に示す 3 原理が含まれると考えられる [3]。その中で矛盾律が問題である。なぜなら前提知識である言語データベースには矛盾する命題が含まれることが多いので矛盾律を単純に適用すると任意の命題が推論可能となってしまう [4]。

表 1 に示す 3 原理のうち排中律を除いた命題論理を直観主義論理、矛盾律と排中律を除いた命題論理を最小論理と呼ぶ。また矛盾律のみを除いた命題論理を排中律論理（仮称）と呼ぶと【表 2】のような 4 種類の論理体系ができる。

表 2　4 種類の論理体系

論理体系	同一律	矛盾律	排中律
古典論理	○	○	○
直観主義論理	○	○	×
最小論理	○	×	×
排中律論理	○	×	○

これら 4 種類の論理体系が実質的に異なることは 3 値論理 [5] などを用いて証明される [2][3]。

④ 矛盾律の取り扱い

矛盾律の問題は LLM より前の知識管理データベースの時代から問題視されており、例えば [4] では、ラフ集合論、様相論理、多値論理を基礎として粒状推論を用いて矛盾を含むデータベースに対処している。

LLM において矛盾律がどのように扱われるかを対話実験により探ってみよう。2025 年 3 月に行った筆者と ChatGPT（無料版）との対話を以下に示す。

3　人間が、古典命題論理を用いて推論した結果が言語データベースに含まれていると仮定すると、そこから学習した LLM も古典命題論理を用いると考えられる。

4　例えば 2025 年 4 月に Yahoo 検索で“トランプは良い”を検索したところ 162,000 件ヒットした。また“トランプは良くない”を検索したところ 33,400 件ヒットした。これらの命題に矛盾律を適用すると任意の命題（例えば「私は宇宙の創造主である」）が推論されてしまう。

5　「真」「偽」の他に第 3 の真理値（仮に「中」と名付ける）をとる論理を 3 値論理という。「中」の扱い方で 3 値論理は複数存在する。

（対話始）

江原　ChatGPT さんは、矛盾律を用いた推論について、どう考えますか。ここで矛盾律とは任意の命題 A と B に対して、「A かつ非 A ならば B」が成り立つことです。

GPT　矛盾律（principle of explosion または ex falso quodlibet）に基づく推論についてですね。これは、古典論理において「矛盾が生じると、どんな命題でも導き出せる」という性質を表しています。つまり、ある命題（A）とその否定（¬A）の両方が真である場合、任意の命題（B）が導かれてしまうというものです。

矛盾律に基づく推論の問題点

この原理を受け入れると、矛盾が一つでも生じた瞬間に、論理体系全体が崩壊してしまう危険性があります。たとえば、もしある数学的体系で（P）と（¬P）の両方が証明されてしまったら、そこから「1 ＝ 2」や「私は宇宙の創造主である」など、無関係な命題まで証明できてしまうことになります。これを防ぐために、通常の論理体系では矛盾が発生しないよう厳格に制約を設けています。

矛盾許容論理（パラコンシステント論理）

しかし、一部の論理体系（paraconsistent logic）では、矛盾があっても任意の命題を導出することを防ぐような仕組みが導入されています。こうした体系では、「矛盾は矛盾として認識するが、それだけでは論理全体が崩壊しない」ように設計されています。たとえば、直観主義論理やダイアルネティズム（矛盾を許容する立場）では、矛盾がある場合の推論規則を再構築することで、爆発を防ぐ試みがなされています。

実用的な視点

・数学・形式論理：矛盾律をそのまま受け入れると、論理体系が破綻するため、通常は矛盾が生じないよう厳密に定義される。

・哲学・認識論：パラコンシステント論理は、実際の思考や人間の認識が必ずしも完全に整合的でないことを考慮するために有用。

・コンピュータサイエンス：非整合なデータベースの処理や AI の推論において、パラコンシステント論理が研究されている。

結論として、古典論理において矛盾律をそのまま使うと危険だが、適切に制限を加えることで、実用的な論理体系を構築することが可能ということになります。矛盾をどう扱うかは、どの論理体系を採用するか次第ですね。

江原　そうですか。それでは矛盾律を用いて「トランプ大統領の言葉は正しい」を証明してもらえますか。

GPT　矛盾律（爆発原理）を用いれば、どんな命題でも証明できてしまうので、「トランプ大統領の言葉は正しい」も形式的には証明可能です。やってみましょう。

証明手順（古典論理に基づく）

1. 矛盾を仮定する

ある命題（P）が真であり、かつ偽であると仮定します。

つまり、

$P \land \neg P$

が成り立つとします。

2. 爆発原理を適用する

矛盾律（爆発原理）より、矛盾が生じると任意の命題

（Q）が導かれます。

$P \land \neg P \Rightarrow Q$

ここで、（Q）を「トランプ大統領の言葉は正しい」と設定すれば、

$P \land \neg P \Rightarrow$「トランプ大統領の言葉は正しい」

が成り立ちます。

3. 結論

したがって、矛盾があるならば「トランプ大統領の言葉は正しい」が証明されました。

重要な注意点

この証明は論理的な意味で正しいですが、現実世界に適用するには問題があります。矛盾律を使うと、任意の命題が証明できるため、「トランプ大統領の言葉は正しい」だけでなく、「トランプ大統領の言葉は間違っている」も証明できてしまいます。つまり、論理体系自体が崩壊しており、この証明には意味がないとも言えます。

結論：矛盾律を使うと何でも証明できてしまうため、厳密な論理的議論では矛盾を含まない体系

を採用するのが一般的です。

(対話終)

このように若干の幻覚はあるものの、ChatGPT(無料版) は現時点では矛盾律の適用に対して常識的な考えを持っているようである。その理由は人間側が矛盾律を用いた推論結果を言語表現としてデータ化することが少ないためと思われる。現時点では人間が作り出した言語表現から LLM が学習されているため、このような常識的な考えに至っているのであろう。しかし今後 LLM 同士の対話データあるいは、LLM が生成した合成データがネット上に増えるにつれて考え方も変わる可能性がある。それより以前に人間側が虚偽の言説 (フェイクニュース) をネット上に載せることが多くなると、LLM が虚偽の情報に基づいて学習してしまう可能性がある。

5 あとがき

LLM を用いた言語生成における幻覚について考察した。特に推論を用いた生成における矛盾律の扱いについて対話実験に基づいて探った。

LLM の学習元である言語データベースに含まれる矛盾する命題に単純に矛盾律を適用すると、任意の命題が推論結果として出力されてしまう。現在の ChatGPT(無料版) は、矛盾率に対して常識的な考えを持っているようであるが、今後の進化しだいでは、矛盾に対する扱い方が変わってしまう可能性がある。万が一にも「矛盾に対処するには戦争が最良の解決手段である」という考えを持つことがないように育ってもらいたいものである。あるいは育ててゆく必要があると考える。

参考文献

[1] Sourav Banerjee, Ayushi Agarwal, Saloni Singla. 2024. LLMs Will Always Hallucinate, and We Need to Live With This, arXiv:2409.05746.

[2] 前原昭二 . 1967. 記号論理入門 , 日本評論社 .

[3] 江原暉将 . 1985. 記号論理学 . 第 26 回数理工学研究会シンポジウム資料 .

[4] 中山陽太郎 . 2020. 矛盾を許容する知識管理データベース , UNISYS TECHNOLOGY REVIEW, 第 145 号 , pp.69-84.