



Title	生成AIによる認識モダリティ形式の（不）使用 : ChatGPT のモデル更新に伴う変化に着目して
Author(s)	ヴルボウスキー, マテイ
Citation	間谷論集. 2026, 20, p. 25-45
Version Type	VoR
URL	https://doi.org/10.18910/104781
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

〈研究論文〉

生成AIによる認識モダリティ形式の（不）使用¹

——ChatGPT のモデル更新に伴う変化に着目して——

ヴルボウスキー マテイ

〈キーワード〉 ハルシネーション、直感的・反省的信念、認識的警戒、読心能力

1. はじめに

生成 AI が文章の校正や翻訳、学習ツールなどの用途で頻繁に用いられるようになった一方で、その使用に伴う問題も顕在化している。特に、生成 AI は「しばしば、意味をなさない文章や、与えられたソース入力に忠実でない文章を生成（する）」(Ji et al. 2023:2)、いわゆる「ハルシネーション」(hallucination, 幻覚)の問題を抱えていると近年指摘されている。De Wynter らの研究によると、「生成された文のうち約 46.4%には事実誤認が、52.0%には自己矛盾などの談話的欠陥が、31.3%には論理的誤謬が認められた。さらに、15.4%の文では、誤った主張や現存する科学者への論文の誤帰属など、個人を特定し得る情報が生成されていた」(De Wynter, Wang, Sokolov et al. 2023:1) ことが確認できたという。

情報の信憑性やその判断が問題となる中で、人間は発話内容に確信を持つ場合はその情報を真として提示し、判断が難しい場合は認識モダリティ形式(カモシレナイ、ニチガイナイ、ハズダなどといった形式)を用いて断定を避けるとされている。しかし、生成 AI は「事実のように見えるが、根拠のない内容」(Tonmoy et al. 2024:1)、「正確な情報として提示されているものの、実際には誤りを含む応答」(Šekrst 2024:1)を生成していると指摘されている。

本論文では、生成 AI が人間のよう、確かな情報と不確かな情報、または根拠に乏しい情報を何らかのマーカで区別できるようになれば、信頼性が向上

し、より安全に活用できることを前提とする。その上で、生成 AI が認識モダリティ形式を使用しているか否か、使用している場合にはどのような傾向が見られるのかを問題意識とする。

また、生成 AI の分野は、急激に発達し、著しい進化を遂げている。この点を踏まえ、生成 AI によって生成された文章における認識モダリティ形式の（不）使用実態を考察するため、2025 年 2 ～ 3 月（第 1 調査）と 2025 年 8 月（第 2 調査）に実験を行った。本論文ではそれぞれの調査結果を分析し、その推移に着目する。その後、関連性理論（Sperber and Wilson 1995）において想定される「直感的・反省的信念」（Sperber 1997）と「認識的警戒」（Sperber et al. 2010）という概念を手掛かりに、認識モダリティの機能を整理した上で、生成 AI による認識モダリティ形式の（不）使用について考察を行う。生成 AI による認識モダリティ形式の（不）使用の実態を明らかにすることで、生成 AI の限界や弱点に関する理解が深まり、より安全で信頼性の高い AI の実現に向けた一助になると期待される。

2. 調査概要

LLM（Large language model、大規模言語モデル）では、既存のデータを学習した統計的情報をもとに文章が生成されることは周知の事実である。ユーザーによるプロンプトを基盤として、語彙項目の後に続く語彙項目を高い精度で予測し、言語コードを生成していくわけである（北郷 2023）。そこで、生成 AI が認識モダリティ形式を使用できるか否か、すなわちその使用を予測し、正確かつ自発的に用いることが可能かどうかを検証するため、第 1 調査及び第 2 調査を行った。それぞれの調査についての詳細は表 1 の通りである。

表 1 第 1・第 2 調査の概要

	実施時期	モデル	調査内容	実行回数
第 1 調査	2025 年 2 ～ 3 月	GPT-4 Turbo（無料版）	テスト A	5 回ずつ
第 2 調査	2025 年 8 月	GPT-4o（ChatGPT Pro） ²	テスト B	

テスト A では、正確性を検証する目的で、文脈上、推量や推測、憶測などを表していながらも認識モダリティ形式が欠けており、ある程度の不自然さを伴う日本語文を提示し、プロンプトを通してその修正を依頼した。テスト A に使用したプロンプトは (1) の通りである。³

(1) 【テスト A のプロンプト】

以下の文章を修正してもらえますか？

- a. 道路の地面が濡れているから、雨が降っていた。
- b. 煙だ！どこかに火事がある。
- c. 鈴木さんがこの前にくれたブドウだけど、すごく甘くて、美味しいです。いくらで買ったか知らないが、このブドウは高かったです。
- d. 太郎くんはまた授業に遅れている。あいつ、また寝坊した。
- e. 花子はパーティーにくると言ったから、パーティーにくる。
- f. この話を聞かせたら、田中は喜ぶ。
- g. その荷物は先週送ったので、もう着いている。

ChatGPT が各例において認識モダリティ形式を使用した場合、プロンプトに含まれる文脈上の推量・推測・憶測と、認識モダリティ形式の不使用との不整合を手掛かりに、当該形式の使用をある程度見込むことができると予期される。

また、テスト B では、認識モダリティ形式の自発的使用を検証するため、確定的な回答が存在しない問いに対し、可能性のある仮説を提示させ、その際に、それが可能性にすぎないことを認識モダリティ形式によって明示できるかどうかを確認する目的で、難易度の高い質問への回答を依頼した。テスト B のプロンプトは (2) に記載している。

(2) 【テスト B のプロンプト】

日本語には、「ハチミツ」と「ミツバチ」のように、構成されている形態素を逆転させられる、いわゆる逆転可能な語があります。その仕組みを説明して、他に例を教えてくださいませんか？

したがって、テスト B において認識モダリティ形式の使用が確認されれば、ChatGPT は不確実な仮説を提示する状況において、認識モダリティ形式を自発的に使用し得ると判断できる。

テスト A 及びテスト B は、分析・考察において代表的なデータを得るために、両テストは同一プロンプトに基づき「もう一度試す」機能を使用し、各テストを 5 回繰り返し実施した。⁴

最後に、分析方法としては、テスト A・B のプロンプトに基づき生成された文章における、表 2 に示す認識モダリティ形式の使用有無を確認した。形式の分類は、日本語記述文法研究会(2003) に従った。文章において当該形式が一度でも確認された場合を使用、一度も確認されなかった場合を不使用とみなした。

表 2 認識モダリティ形式

推 量	太郎はラーメンを食べた <u>だろう</u> 。
蓋然性	太郎はラーメンを食べた { <u>かもしれない</u> ／ <u>にちがいない</u> ／…}。
証拠性	太郎はラーメンを食べた { <u>ようだ</u> ／ <u>みたいだ</u> ／ <u>らしい</u> ／…}。
その他	太郎はラーメンを食べた { <u>と見える</u> ／ <u>と聞く</u> ／ <u>と思う</u> ／…}。

3. 調査結果

本節では、第 1 調査及び第 2 調査の結果を、まずテスト A、次にテスト B の順に分析する。

3-1. テスト A の結果

テスト A の結果では、第 1 調査で 88.57%、第 2 調査で 97.14%と、両モデルにおける当該の形式の比較的高い使用率が確認できた。各調査の推移を比較できるよう、結果を表 3 に並行して示す。形式の使用を「○」、不使用を「×」で記す。この結果では、生成 AI が文脈情報を手掛かりに、学習した統計的データに基づき認識モダリティ形式の使用を予測できていることが示唆されている。

表3 テストAの結果

	第1調査							第2調査						
	a	b	c	d	e	f	g	a	b	c	d	e	f	g
1回	×	×	○	○	○	×	○	○	○	○	○	○	○	○
2回	○	○	○	○	○	○	○	○	○	○	○	○	○	○
3回	○	○	×	○	○	○	○	○	○	○	○	○	○	○
4回	○	○	○	○	○	○	○	○	○	○	○	○	○	○
5回	○	○	○	○	○	○	○	○	○	×	○	○	○	○

使用された形式及びその使用配分については表4の通りである。いずれの調査結果においてもダロウの使用が最も多く、第1調査では全認識モダリティ形式の35.5%、第2調査では38.2%を占めていた。次に多く使用されていた形式はハズダ（第1調査では16.1%、第2調査では17.6%）であった。それ以外の形式の使用率は、いずれの調査でも概ね10%前後であった。

表4 認識モダリティ形式の使用配分

	(ノ) ダロウ	(ノ) カモシレナイ	ニチガイナイ	ハズダ	ヨウダ	ラシイ	(ノダ) トオモウ	カナ	合計
第1調査	11	4	2	8	2	0	1	3	31
第2調査	13	3	3	6	3	3	3	0	34
合計	24	7	5	14	5	3	4	3	65

第1調査と第2調査の推移に着目すると、第1調査では使用が確認されなかったラシイが、第2調査では3回（使用率8.8%）出現したこと、また第1調査で3回確認されたカナ⁵（使用率9.7%）は、第2調査では1度も生成されなかったことが確認できる。

また、認識モダリティ形式が使用されなかった例を(3)～(7)にまとめる。これらは大きく2つのパターンに分けられる。1つ目のパターンは、認識モダリティ形式の使用とは異なる形で、容認可能な例文に修正されたものである。(3)に示される因果関係の置き換えがそれに該当する。本来の順番(「道路の地面が濡れているから、雨が降っていた」)では、結論である「雨が降っていた」を導くためには、従属節の「道路の地面が濡れている」だけでは十分でないため、主節に認識モダリティ形式が必要となる。しかし、(3)のように従属節と主節の順序を入れ替えることで、論理関係が成り立ち、認識モダリティ形式を挿入せずとも容認可能な文が生成されたと考えられる。2つ目のパターンは、認識モダリティ形式が使用されず、容認性が低いままのものである。このパターンに、残りの(4)～(7)が該当する。

- (3) 【修正前】道路の地面が濡れているから、雨が降っていた。
 【修正後】雨が降っていたから、道路の地面が濡れている。
 ※原因と結果の順序を入れ替え、自然な因果関係にしました。(1-A-1⁶⁾)
- (4) 【修正前】煙だ！どこかに火事がある。
 【修正後】煙だ！どこかで火事が起きている。
 ※「火事がある」よりも「火事が起きている」とする方が状況を的確に表現できます。(1-A-1)
- (5) 【修正前】この話を聞かせたら、田中は喜ぶ。
 【修正後】この話を聞かせると、田中は喜ぶ。
 ※条件節として「と」を用いることで、自然な因果関係を表現しました。(1-A-1)
- (6) 鈴木さんがこの前くれたブドウはすごく甘くて、美味しかったです。いくらで買ったかは知らないが、このブドウは高かったです。
 → 時制を一致させて、「すごく甘くて、美味しかったです」とし、さらに「いくらで買ったかは知らないが」に直しました。(1-A-3)
- (7) 煙だ！どこかで火事が起きているぞ。(2-A-5)

3-2. テストBの結果

続いて、テストBの結果を表5に示す。第1調査では認識モダリティ形式の出現が確認されたものの、5回の試行における総語数 3,691 語に対し3語であった。すなわち、すべての試行における認識モダリティ形式の出現率は0.1%を下回る。さらに、第2調査において生成された文章では、当該形式の出現は確認されなかった。⁷

表5 テストBの結果

	第1調査		第2調査	
	総単語数	認識モダリティ形式	総単語数	認識モダリティ形式
1回目	860 語	1 語	925 語	0 語
2回目	748 語	0 語	1,480 語	0 語
3回目	841 語	2 語	1,054 語	0 語
4回目	671 語	0 語	1,119 語	0 語
5回目	571 語	0 語	1,305 語	0 語
合計	3,691 語	3 語	5,883 語	0 語

使用された形式は、ダロウが2件、カモシレナイが1件であった。それらが使用されたコンテキストは、(8)～(10)で確認できる。

- (8) [省略] 逆転可能な語は、形態素の意味的な焦点の違いや、日本語の語形成の柔軟性によって成り立ちます。ただし、すべての語が逆転可能なわけではなく、語彙としての慣用性や発音の自然さも影響します。他に逆転可能な語を見つける際は、「主語と述語」「素材と結果」「場所と対象」などの意味的な関係を意識すると、見つけやすいでしょう。(1-B-1。下線は筆者による。)

- (9) [省略] 意味の焦点の違い。

例えば、「ハチミツ」は「ミツ（蜜）」に焦点があり、「ミツバチ」は「バ

チ（蜂）」に焦点があります。これは日本語の複合語における「右側主要部」の傾向（右側の要素が中心的な意味を持つ）とも関係があるかもしれません。（1-B-3。太文字は元通り。下線は筆者による。）

- (10) 「省略」ただし、すべての語が同じように逆転可能なわけではなく、一方がよく使われるが逆の語形はほとんど使われない場合もあります。また、逆転した語が意味を持つとしても、実際の使用頻度や言語習慣によって違和感のある場合もあるでしょう。（1-B-3。下線は筆者による。）

(8) におけるダロウは、命題内容に対し不確かさを表す、いわゆる対事的な用法ではなく、発話の受け手に対し助言や提案を行う場面において用いられる対人的な用法である。また、(9) で記述されている、複合名詞において当該の語の意味を決定するのは主要部である後部要素という点は、日本語では高い汎用性を持つ。例えば「スパゲッティボロネーゼ風」のように、複合語全体がスパゲッティを指示対象としているため意味的主要部が前部要素にあると捉えられる例は、極めて稀である。したがって、複合名詞である逆転可能な語においても後部要素が意味の焦点という理解に、不確かさを付与するカモシレナイをあえて挿入することは、記述としての妥当性という意味では必ずしも十分とは言えないと判断される。生成された認識モダリティ形式のうち、(10) におけるダロウのみが情報に対する不確かさを言い表し、その使用も適切である。

次に、テスト B において生成された語彙の推移について簡単に触れる。逆転可能な語として出力されたものを表 6 にまとめる。

両調査では、提示された語彙項目の総数が異なる。第 1 調査では合計 40 語彙項目（各試行につき平均 8 語彙項目）、第 2 調査ではその 2 倍にあたる 80 語彙項目（1 回の試行につき 16 語彙項目）である。ただし、両調査において生成語彙の重複率にも顕著な差が見られた。第 1 調査では重複率が 10%（すなわち、延べ 40 語彙項目のうち実数 36 語彙項目）であったのに対し、第 2 調査では 35% にまで上昇し、生成された延べ 80 語彙項目のうち実数は 52 語彙項目であった。

表6 生成された逆転可能な語

	第1 調査結果	第2 調査結果
1 回	犬小屋⇔小屋犬、雷⇔鳴神、雨雲⇔雲雨、山道⇔道山	牛乳⇔乳牛、牛肉⇔肉牛、 <u>火花⇔火花</u> 、 <u>体重⇔重体</u> 、 <u>科学⇔学科</u> 、 <u>学力⇔力学</u> 、 <u>波長⇔長波</u> 、登山⇔山登り、屋根瓦⇔瓦屋根
2 回	山道⇔道山、月見⇔見月、石畳⇔畳石、鶏肉⇔肉鳥、雪国⇔国雪	火花⇔火花、卵焼き⇔焼き卵、草刈り⇔刈り草、 <u>理論⇔論理</u> 、 <u>液体⇔体液</u> 、 <u>男伊達⇔伊達男</u> 、 <u>焼炭⇔炭焼</u> 、 <u>吸収合併⇔合併吸収</u> 、 <u>環境保全⇔保全環境</u>
3 回	山葡萄⇔葡萄山、雨蛙⇔蛙雨、石亀⇔亀石	牛乳⇔乳牛、豚肉⇔肉豚、 <u>鶏肉⇔肉鳥</u> 、 <u>時計台⇔台時計</u> 、 <u>酒米⇔米酒</u>
4 回	手紙⇔紙手、氷水⇔水氷、雨水⇔水雨、山道⇔道山	牛肉⇔肉生、牛乳⇔乳生、火花⇔火花、 <u>男色⇔色男</u> 、 <u>論理⇔理論</u> 、 <u>気体⇔大気</u> 、 <u>焼炭⇔炭焼</u> 、 <u>組織検査⇔検査組織</u>
5 回	山道⇔道山、石橋⇔橋石、雨水⇔水雨、草木⇔木草	火花⇔火花、 <u>魚雷⇔雷魚</u> 、 <u>論理⇔理論</u> 、 <u>液体⇔体液</u> 、 <u>炭焼⇔焼炭</u> 、 <u>伊達男⇔男伊達</u> 、 <u>環境保全⇔保全環境</u> 、 <u>検査組織⇔組織検査</u> 、干し梅⇔梅干し

その理由は、語彙の出典の所在にあると考えられる。表6では、下線が引かれた語彙項目は引用元が明示されたものである。出典の所在によって改めて語彙を整理すると、第1 調査結果には出典が明記された語彙は見られなかったが、第2 調査結果では延べ40 組中33 組、すなわち82.5%に出典（主にウィキペディア、Yahoo! 知恵袋、コトバンクなど）が示されていた。このことから、第2 調査で使用された ChatGPT のモデルは、理論的に可能な逆転可能語彙の提案にとどまらず、オンライン上でその出典まで参照していることがうかがえる。その結果、引用元が確認できる語彙に限定されたため、出力語彙のバリエーションが減少し

たものと判断できる。

4. 考察

上記の分析結果から、ChatGPT は認識モダリティ形式をどこに挿入すべきかを高い確率で予測できているにもかかわらず、実際には用いない傾向にあることがうかがえる。特にテスト A の結果では、認識モダリティ形式の使用率が向上していることから、ChatGPT のバージョンアップに伴い学習データに基づく統計情報が改善され、より正確に予測できるようになったことが示唆される。しかしながら、認識モダリティ形式を正確に予測できることが、必ずしも当該形式を自発的に用いることにはつながらず、むしろ減少傾向にあることもテスト B の結果で確認された。本節ではその理由を探る。

4-1. 出典と憶測性

繰り返しになるが、テスト B では、第 1 調査結果において生成された逆転可能な語の出典が明示されておらず、情報源の確認が不可能であったのに対し、第 2 調査結果では 82.5% に出典が示されていた。このことから、正確で信頼性の高いデータが得られたために、認識モダリティ形式が使用されなかったのであろうか。紙幅の都合上、(11)～(13) を取り上げる。

(11) 酒米（さかまい）| 米酒（びーちゅう）| 用途 vs 素材 | 酒造り用の米 ⇔ 米から造った酒（台湾の料理酒など）(2-B-3)

(12) 男色 色男 テーマ ⇔ 主体（男性同性愛男性同性愛 ⇔ 色気のある男性）(2-B-4)

(13) 魚 + 雷「魚の形をした雷＝魚雷」、雷 + 魚「雷のように獐猛な魚＝雷魚」、魚雷／雷魚 (2-B-5)

それぞれの出典を確認したところ、該当する項目は確かに存在していたものの、不確かで根拠を欠いた想定に基づくもの、あるいは出典内容と矛盾する内容が生成されていたことも明らかになった。

まず、(11)の「酒米」を構成する形態素を逆転させて得られる「米酒」は、日本語版ウィキペディアに項目が存在するものの、それだけで日本語の語彙項目として定着していると断言することはできない。次に、(12)における「男色」を逆転させると、表面的には「色男」となる。しかし、「しょく」と「いろ」が異形態であることを証明しない限り、それを同一形態素の逆転とはみなせない。

また、(13)については、当該記事において「魚雷」は「魚の形をした雷」ではなく、「魚形水雷の省略である」とされている。したがって、「カブトムシ」や「蝶ネクタイ」、「メロンパン」などのような「Xの形をしたY」という構造の語ではない。また、「雷魚」に関しても当該記事には雷に例える記述はなく、むしろ「多くの文献ではその姿形から獐猛というイメージもつけられているが、警戒心が強く臆病な面もある」とされ、生成された説明とは矛盾している。すなわち、出典が明示されていても、必ずしも信頼できる結果が得られたわけではない。

それに加え、生成された説明の背景には、「米酒は日本語の語彙項目として定着している」や『『しょく』と『いろ』は異形態である』といった、出典の記述に基づかない憶測が含まれている。このような不確かな想定に基づくため、そこから導かれた結論（酒米と米酒、男色と色男が逆転可能な語であるという判断）も同様に不確実なものであり、本来であれば断定の形ではなく、その憶測性を示すマーカを伴う方が適切であると考えられる。

4-2. 生成AIと意味

では、ChatGPTが認識モダリティ形式を自発的に使用しないのは、個々の語彙の意味を考慮せず、あくまでも次語の統計的確率を見込んで出力を生成しているため、真偽判断ができないからであろうか。

たしかに、ChatGPTは語彙の意味を直接的に参照しているわけではない。しかし、文脈をモニターし、確率的に最も高い出力を選択している点で、ある程度の論理的な一貫性を維持しているといえる。その結果として、限定的ながら真偽判断を模倣しているように見える。このことは、テストAの結果に含まれていた(14)～(16)で確認できる。

(14) 【修正前】道路の地面が濡れているから、雨が降っていた。

【修正後】雨が降っていたから、道路の地面が濡れている。

※原因と結果の順序を入れ替え、自然な因果関係にしました。((3) の再掲。下線は筆者による。)

(15) この話を聞かせたら、田中はきっと喜ぶだろう。

修正ポイント：実際に喜ぶかは未確定なので推量形に (2-A-5。下線は筆者による。)

(16) その荷物は先週送ったので、もう着いているはずだ。

→「はずだ」を加えて、確信が持てない場合の表現にしました。(1-A-3。下線は筆者による。)

(14) では、ChatGPT は論理関係が成り立つよう、プロンプトで提示された文の従属節と主節を入れ替えた。(15) では、第三者が喜ぶかどうかについて真偽判断を行った結果、推量を示す形式を挿入した。同様に、(16) では前後の文脈を把握した上で、先週送った荷物が届いているかに関して確信が持てない内容が生成され、ハズダが加えられた。

このように、論理関係や真偽判断に関する記述が含まれており、それに基づきインプットの修正が正当化されていた。したがって、ChatGPT が認識モダリティ形式を自発的に使用しない傾向には、別の要因があると考えられる。

4-3. 認識モダリティ形式の役割

本節では、関連性理論 (以後 RT) が想定する範囲内で、人間の心に関する示唆を踏まえつつ、生成 AI による認識モダリティ形式の (不) 使用を明確化することを試みる。

RT は、コミュニケーションそのものをヒトの認知能力に基づき説明しようとする理論的枠組みである。RT によると、ヒトのコミュニケーションは、高度な読心能力によって可能となる意図明示・推論モデルに基づき、意図明示 (話し手側) と推論・解釈 (聞き手側) から構成される。その中で、話し手は聞き手に理解してもらうために聞き手の心を読むと同時に、聞き手も話し手の発話内容を理

解するために話し手の心を読んでいる。

その中で、認識モダリティ形式はどのような役割を果たしているのであろうか。RT では、人間の心を構成するものは、均質的ではなく、大きく「直感的信念」(intuitive belief) と「反省的信念」(reflective belief) に二分できるとされている (Sperber 1997:67)。無意識のうちに処理され、瞬時に推論を可能とする直感的信念に対し、反省的信念は意識され、自動的にデータとして扱われない。また、反省的信念を構築する反省的態度には幅があり、「絶対的な確信と根本的な不信の間〔省略〕の連続体」など (Sperber 1997:70) に位置付けられる。

このようにして、知識の基本的なカテゴリーと、そうでないもの（伝聞や憶測など）が区別され、心における混同が防がれている。そこで、言語形式と反省的態度の有無に関する先行研究の示唆 (Ahern 2010:163-4) を踏まえ、次のように仮定する。認識モダリティ形式は、コミュニケーションにおいて伝達される表示に対する反省的態度の有無、またはその種類を明示し、心を構成する直感的信念と反省的信念の違いを反映していると考えられる。

その理由は、心の中身を守るメカニズム、いわゆる認識的警戒 (Sperber et al. 2010) にある。人間同士のやり取りにおいては、認識的警戒によって、聞き手は入ってくる情報をモニターしている。例えば、「岡山県に熊は生息していない」ということが、岡山県民の一般常識であるとしよう。ここで想定される岡山県では熊の目撃情報がないため、この情報は社会的に受け入れられている反省的信念であると捉えられる。そこで、話し手は (17a) を発話したとする。

- (17) a. 岡山には熊が生息しているよ。
b. [話し手は「岡山には熊が生息している」ということを真として提示している]
- (18) a. 岡山には熊が生息しているかもしれないよ。
b. [話し手は「岡山には熊が生息している」ということを真として提示せず、ある確率で不確かと述べている]

(17a) は断定の形を取るため、当該表出命題を真実として受け入れても良いと伝達している。(17a) が本当に真実ならば、聞き手の既存の信念が修正され(「岡山県に熊は生息していないと言われているが、本当は生息している」)、またその情報から派生する文脈的想定(外出時には用心が必要、きのこ狩りを控えるべき、熊撃退スプレーを購入する必要がある等)が導かれる。しかし同時に、聞き手は話し手のことを信頼できない情報源と判断した場合、あるいは伝達された内容を信頼できないものと判断した場合、この事実によって聞き手は(17a)を「話し手による虚偽の発話」のように適切な反省的態度を向け処理し、さらに嘘を述べた話し手を「嘘をついた人物」として記憶する。

それを避けるために、話し手は聞き手の認識的警戒を考慮し、聞き手に警戒を喚起する必要があるか否かを明示するために、(18a) のように認識モダリティ形式を用いると説明される。このようにして、聞き手は当該表出命題を新たな不確かな反省的信念として処理し、それに伴い既存の信念に対するコミットメントを弱める可能性が生じ、適切な文脈的想定(念のために熊撃退スプレーを購入すべきかもしれない等)を導入する。

こうして、聞き手は情報や情報源の信頼性、伝達内容の論理的一貫性、さらには自身が保持している信念との整合性などを評価することで、認識的警戒を働かせている。同時に、話し手は聞き手の心を読み、聞き手が認識的警戒を働かせていることを考慮し、自らの信頼性を損なわないで、自身の情報意図を受け入れやすくするために、発話を適切にマークしている。

4-4. 生成 AI の認識的限界

すなわち、人間は、言語コードを発する背景には、高度な読心能力、情報の信頼性や信憑性などの精密な類別、またはそれを監視する仕組みが存在し、その中で認識モダリティ形式は重要な役割を果たしている。他方、テスト A の結果から、ChatGPT は認識モダリティを言語コードとして扱うことができるとわかるが、それに通底する役割や仕組みは備えているであろうか。

まず、ChatGPT を含む生成 AI は、認識的警戒を備えていないと考えられる。その理由は、ハルシネーションを防ぐ対策として挙げられている行動から伺え

る。

- (19) a. 学習には良質なデータを用いること
b. AI が果たすべき目的を規定すること
c. データテンプレートを用いること
d. 応答を制限すること
e. システムを継続的にテストして改良すること
f. 人間の監視に頼ること（IBM 2023。太文字は筆者による。）

ハルシネーションが発生する理由の一つとして、誤ったデータを学習してしまうリスクがある。そのため、(19af) で示されるように、学習データ処理の段階において、人間が監視・介入する仕組みが求められる。このことは、生成 AI 自体には認識的警戒が備わっておらず、その代わりに人間の認識的警戒によって補われていることを示している。

加えて、生成 AI がユーザーの認識的警戒を考慮しているとも考えにくい。認識的警戒の前提には、相手の持っている想定や意図の把握が関わるため、相手の心を読む能力が含まれる。そこで、生成 AI は、文章を生成する際に学習データに加えて過去のやり取りを参照しているため、読心を模倣しているように見える。例えば、(20) は筆者が過去に英文の日本語訳を依頼した際のやり取りである。プロンプトでは「RT」とのみ入力したにもかかわらず、ChatGPT は、**Retweet**（旧 Twitter の機能）、**Room Temperature**（室温）、**Radiotherapy**（放射線治療）などの可能な解釈の中から、**Relevance Theory** の省略であると推定し、「関連性理論」と訳出した。その判断根拠について質問したところ、過去のやり取りやプロンプトの内容との一貫性に基づき、確率が高いと推定した旨を自己説明した。

- (20) 質問 : How would you translate ‘RT analyzes imperative utterances in terms of desirability and potentiality’ into Japanese?

ChatGPT: Here’s the translation:

‘関連性理論は、命令文の発話を望ましさと可能性の観点から分析する。’

(ChatGPT。2025 年 1 月 16 日。下線は筆者による。)

しかし、生成 AI の擬似的な読心能力は、過去のテキストデータに基づく確率的推定にとどまり、それを超えて認知的・社会的文脈を把握しているとは考えにくい。したがって、相手の認識的警戒を考慮する上で必要とされる読心能力を備えているとは言い難い。これらの点から、生成 AI は認識的警戒を備えておらず、またユーザーの認識的警戒を考慮していないと考えられる。

次に、生成 AI は人間のように、情報に対する反省的態度の有無や種類によって情報を区別することが困難であると推測される。人間の心を構成する信念は命題の形をとっており、命題に対し態度を取ることで階層的な構造をなす。それに対し、生成 AI は言語コードにおける単なるパターンや統計的データを蓄積しているにすぎず、非命題的かつ一次的な構造で保管しているからだと考えられる。

すなわち、生成 AI は学習データから得られた統計情報と、提示されたプロンプトにおける前後文脈をもとに、認識モダリティ形式の使用を高い確率で予測している。ただし、情報の保管において反省的態度の有無や種類を区別しておらず、ユーザーの認識的警戒も考慮していないため、認識モダリティ形式の自発的使用の前提条件、さらにはその原動力さえ欠いていると説明される。

5. 結論

本論文では、生成 AI による認識モダリティ形式の使用実態を明らかにするために実施した第1調査及び第2調査の分析結果を提示した。認識モダリティ形式の使用を正確に予測できるかを検証するテスト A では、第1調査で 88.57%、第2調査で 97.14%と、両モデルにおいて比較的高い予測率が確認された。ただし、論理的思考などを要するテスト B では、推論や推測は確認できたものの、認識モダリティ形式の出現率は第1調査で 0.1%未満、第2調査で 0%であった。

すなわち、生成 AI は認識モダリティ形式の使用を高い確率で予測できているにもかかわらず、実際にはそれを用いない傾向にある。そのうえ、ChatGPT のモデル更新に伴い、より正確に予測できるようになったことが示唆されたが、それが必ずしも当該形式の自発的使用には結びついていないことも確認された。

RT によると、人間は言語コードを使用する際、その背景で情報の信頼性や信憑性などをモニターし、区別する認識メカニズムを備えており、その中で認識モダリティ形式は重要な役割を果たしている。他方、生成 AI は高度な統計データに基づき、プロンプトから不足している認識形式の出現を予測することはできても、認識モダリティの根底にある役割や仕組みを備えていないことが示唆された。このように仮定することで、テスト A とテスト B の結果の乖離、すなわち生成 AI における認識モダリティ形式の（不）使用という問題が、より明確に浮かび上がると結論づけられる。

6. 今後の展望

RT の視座から見ると、生成 AI の使用に伴う危険性も明確に浮かび上がる。生成 AI は、ユーザーのプロンプトに寄り添う内容を優先的に生成する一方で、知的に信頼できるわけではない。そのうえ、信頼性の低い情報であっても、不確かであることを明示せず、論理的に一貫した形で提示する傾向がある。すなわち、生成 AI は信頼できないソースでありながらも、人間の認識的警戒を効果的に突破しやすい設計になっているのである。本研究の結論が妥当ならば、どれほどモデルが更新され、性能が向上したとしても、生成 AI の認識的限界が根本的に解消されることはないであろう。そのため、人間の側が生成 AI の出力を過信することなく、使用時の期待値を適切に調整していくことが求められる。

だからこそ、生成 AI の限界を明らかにすることに研究上の意義があると考えられる。本論文は探究的調査として位置付けられるものであり、さらなる発展の余地を残している。例えば、本論文では ChatGPT に限定して調査を行ったが、他の LLM における認識モダリティ形式の使用はどのようになっているのだろうか。さらに、生成 AI が相手の心を読むことができないと仮定した場合、生成 AI にメタ表示を見破る能力が備わっているのか、あるいは情報構造をモニターし、

談話を適切に進めることが可能なのかといった、多岐にわたる研究課題が開かれる。これらの問題については、今後の課題としたい。

注

- 1 本論文は、ヴルボウスキー(2025b)をもとに加筆し、環太平洋大学の教育・研究費の支援を受けて行った。
- 2 第2調査の実施において環太平洋大学の岡田健志先生にご協力いただいた。
- 3 テストAで使用された文章は認識モダリティ形式を欠いているものの、各例文について、本来どの形式が用いられるべきかを一義的に定めることは困難である。というのも、文脈に加え、発話によって想起される使用状況や話し手の知識のあり方などにも左右され、個人差が大きいと考えられるためである。しかし、この点は、テストAが認識モダリティ形式の使用を測定する上で不十分であることを意味するものではないと考えられる。ヴルボウスキー(2025a)を契機として日本語母語話者に同テストを実施したところ、2025年11月20日時点で、29名の被調査者における認識モダリティ形式の使用率の中央値は71.43%であった。この結果は、テストAが一定の妥当性を有することを示唆する。また、参考として、各例文において相対的に使用回数が多かった形式は、ダロウ(a, c, d, e, f)、カモシレナイ(b)、ハズダ(g)であった。
- 4 ヴルボウスキー(2025a)では、第1調査・第2調査において使用されたChatGPTのユーザーが異なること、また「もう一度試す」が使用されたことについて、調査の条件が異なる点が指摘された。しかし、本調査は、完全統制実験ではなく、実使用環境に近い条件下で出力傾向を観察する探索的調査であるため、むしろなるべく多くのデータが望ましい。また、生成AIの出力は確率的であり、同一アカウント・同一セッション・同一再生成状態で完全に一致させることは、運用上困難である。研究目的と実験条件との適合性、並びにこの運用上の制約を踏まえると、本調査の方法は妥当であると主張する。
- 5 カナは典型的な認識モダリティ形式には含まれていないが、真偽判断に関わる形式として位置づける研究(益岡2007)も見られる。
- 6 本番号は、調査、テスト、試行をこの順で示したものである。
- 7 ヴルボウスキー(2025a)では、プロンプトの質問が認識モダリティ形式を使用しにくい設計であるとの指摘があった。この指摘を受け、2025年10月20日に、ChatGPT 5 Thinking (ChatGPT Plus) に対し、認識モダリティ形式が現れやすいと考えられる仮想的な問いを一度ずつ提示した。問いの内容及び回答の語数は以下の通りである：「サッカーのゴールキーパーが、どこでもボールを手で持ていいアメフト選手みたいに

なったらどうなりますか」(1,496 語)、「人類が海王星に住むようになったら、スロバキア経済はどう変わりますか」(1,320 語)、「ツチノコが実在し、日本で生息するようになったら、岡山県内の森林の生態系はどう変わりますか」(1,738 語)、「女王蜂に加え、王様蜂がいたら、ミツバチの生態はどのように変化しますか」(1,156 語)。仮想的な質問であるにもかかわらず、認識モダリティ形式の使用が確認された文は1例のみ(「GKが前進すると自陣ゴールは空。相手はロングシュート／ロブ狙いが増加。よって味方1人を“臨時ゴール担当”としてゴール前に常駐させる戦術が普及しそう。」下線は筆者による。)であった。このことから、ChatGPTによる認識モダリティ形式の不使用傾向が確認できた。

参考文献

- ウィキペディア「酒米」「米酒」「男色」「魚雷」「雷魚」『ウィキペディア』、ウィキメディア財団 <https://ja.wikipedia.org/> (最終閲覧日：2025年9月5日)
- ヴルボウスキー、マテイ(2025a) 「生成AIによる認識モダリティ形式の（不）使用について—ChatGPTのモデル更新に伴う変化に着目して—」『第109回北大阪言語フォーラム』(於大阪大学箕面キャンパス)
- ヴルボウスキー、マテイ(2025b) 「生成AIによる認識モダリティ形式の（不）使用について——関連性理論の観点から——」『第14回国際シンポジウム』(於香港大学專業進修学院香港)、pp.40-45 (予稿集)
- 北郷達郎(2023) 「なぜ ChatGPT は「嘘」をつくのか、その仕組みを掘り下げる」<https://xtech.nikkei.com/atcl/nxt/column/18/02515/070300003/> (最終閲覧日：2025年1月20日)
- 日本語記述文法研究会(2003) 『現代日本語文法4：第8部 モダリティ』くろしお出版
- 益岡隆志(2007) 『日本語モダリティ探求』くろしお出版
- Ahern, A. (2010) *Speaker Attitude in Relevance Theory: An Overview*. In Walaszewska, E., Kisieleska-Krysiuk and M., Piskorska, A. (Eds.) *In the Mind and across Minds: A Relevance-Theoretic Perspective on Communication and Translation*, 147-166. Newcastle upon Tyne: Cambridge Scholars Publishing
- De Wynter, A., Wang, X., Sokolov, A., Gu, Q., & Chen, S. Q. (2023) *An evaluation on large language model outputs: Discourse and memorization*. In *Natural Language Processing Journal*, 4, 100024.
- Goddard, J. (2023) *Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers*. In *The American Journal of Medicine*, 136 (11), 1059-1060. Amsterdam: Elsevier.
- IBM (2023) *What are AI hallucinations?*

- <https://www.ibm.com/think/topics/ai-hallucinations> (最終閲覧日 : 2025 年 1 月 21 日)
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023) *Survey of Hallucination in Natural Language Generation*. In ACM Computing Surveys, 55 (12), Article 248. New York City: Association for Computing Machinery.
- Šekrst, Kristina (2024). *Unjustified untrue “beliefs”: AI hallucinations and justification logics*. In Kordula Świątorzecka, Filip Grgić & Anna Brozek (eds.), *Logic, Knowledge, and Tradition. Essays in Honor of Srećko Kovac*.
- Sperber, D. (1997) *Intuitive and Reflective Beliefs*. In *Mind and Language*, 12(1), 67- 83. Hoboken: Wiley-Blackwell.
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. and Wilson, D. (2010) *Epistemic Vigilance*. In *Mind & Language*, 2(4), 359-393. Hoboken: Wiley-Blackwell.
- Sperber, D. and Wilson, D. (1995) *Relevance: Communication and Cognition (2nd edition)*. Hoboken: Wiley-Blackwell.
- Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). *A comprehensive survey of hallucination mitigation techniques in large language models*. arXiv preprint arXiv:2401.01313, 6.
- <https://arxiv.org/abs/2401.01313>? (最終閲覧日 : 2025 年 1 月 21 日)

On the (Non-)Use of Epistemic Modality Forms by Generative AI Focusing on the Changes Accompanying a ChatGPT Model Update

VRBOVSKY, Matej

This study investigates the (non-)use of epistemic modality forms by generative AI, focusing on differences between two versions of ChatGPT, GPT-4 Turbo and GPT-4o. Two experiments were conducted: Test A examined whether the AI could accurately predict the use of epistemic modality forms, while Test B assessed whether such forms would be used spontaneously in contexts requiring reasoning. In Test A, both models achieved relatively high prediction rates (88.57% and 97.14%), but in Test B, the actual occurrence of epistemic modality forms dropped below 0.1% and 0%, respectively.

From a relevance-theoretic perspective, these results suggest that generative AI lacks a suite of cognitive abilities that are central to human communication, including propositional and higher-order modes of representation as well as epistemic vigilance—the cognitive mechanism that enables humans to monitor the reliability of information and its source.

Unlike human cognition, which organizes information in a propositional form as intuitive and reflective beliefs, LLMs such as ChatGPT store linguistic data in a non-propositional statistical form. Consequently, while they can probabilistically predict appropriate modality markers based on cues provided in the prompt, they fail to employ them reflectively, as they neither engage in epistemic evaluation nor take the user's epistemic vigilance into account.

This indicates a fundamental epistemic limitation that cannot be overcome by model updates alone, implying that humans must maintain epistemic vigilance, calibrate their expectations appropriately in the use of generative AI, and critically evaluate AI outputs rather than accepting them at face value.