生成 AI が「もっともらしいけれど実は間違っている情報」を答えてしまう現象（ハルシネーション）は、よく「バグ」や「知識不足」のせいだと誤解されます。

しかし実際は、**受験生がマークシートで“分からなくてもとりあえず塗る”のと同じように、AI も“当てずっぽうで答えた方が得点になる”ように訓練されている**のが原因です。

この構造的な問題に対して、OpenAI は公式ブログで原因・対策・改善の成果を詳しく分析・発信しました。

AI の「ハルシネーション（誤情報）」はなぜ起こるのか

― OpenAI による構造的課題の分析と改善提案 ―

1. はじめに：AI の信頼性に関わる課題とは

近年、生成 AI の利用が急速に拡大する一方で、「**AI がもっともらしく話すが実は間違っている**」という現象 ― いわゆる**ハルシネーション（hallucination）** ― が、ユーザー体験や業務利用において大きな課題となっています。

OpenAI は、こうした現象が偶発的なバグではなく、**AI モデルの学習および**

評価プロセスに起因する“構造的な問題”であると指摘しています。

2. ハルシネーションの構造的原因

❗ 問題の本質：推測を促す評価制度

現在の AI のトレーニングでは、以下のような評価ロジックが主流です：

- 「正解を出す」 → 加点
- 「間違える」 → 減点
- 「分からない」や「答えを控える」 → 加点なし（=実質 0 点）

このロジックにおいて、**AI が「わからない」と正直に答えるよりも、推測でも何かを答えた方がスコアが高くなる**という仕組みになってしまっています。
その結果、モデルはこう学習します：

「間違ってもいいから、何か答えた方が得」

たとえば、「ある人の誕生日を教えてください」という問いに対し、モデルは確証がなくても「9 月 10 日」と答える方が、「わかりません」と答えるより評価される。これが**幻覚的回答の温床**になっているのです。

3. 実例：GPT モデルの回答傾向の比較

OpenAI が行った評価（SimpleQA）によると、以下のような傾向が見られました：

モデル	正答率	誤答率（=幻覚）	無回答率
GPT-4 mini（o4-mini）	24%	**75%**	1%
GPT-5 thinking-mini	22%	**26%**	**52%**

この比較から、**「答えを控える」という姿勢を持つ GPT-5 mini の方が、誤答（幻覚）を大幅に抑制している**ことがわかります。
正答率ではやや劣って見えるものの、**実用上の信頼性ははるかに高い**と評価されています。

4. OpenAI の提案：評価制度の見直し

OpenAI は、モデルそのものだけでなく、**「評価方法」そのものを見直すべきである**と提言しています。

✓ 改善提案の方向性：

- **誤情報には厳しい減点を与える**（自信ありげな誤答を抑止）
- **「わからない」「情報が不足している」といった不確実性の表現には部分点を与える**
- **正答／誤答／無回答のバランスを踏まえたスコア設計**

このような改善により、**"わからないときは無理に答えず、正直に答えるAI"**を育てるインセンティブが生まれます。

5. 今後の展望：より信頼される AI へ

GPT-5 ではこうした方向性に沿った改良が施され、**実際に誤情報の頻度は GPT-4 より大きく減少**しています。

しかし OpenAI は、「この問題を根本から解決するには、**評価制度の構造改革が不可欠**である」と強調しています。
言い換えれば、**AI の"性格"は、どんなテストで褒められ、どんな回答で叱られるかで決まる**ということです。

6. おわりに：実務への示唆

この分析は、業務で生成 AI を導入・評価する際にも重要な視点を提供します：

- **「精度」だけでなく、「誠実さ」「無理な推測をしない姿勢」も評価指標に含めるべき**
- モデルに「誤った自信」を持たせないための設計・トレーニングが重要
- **評価の仕方が AI の行動を左右する**ことを理解する

🔗 参考資料

- OpenAI 公式ブログ

  Why Language Models Hallucinate
- 解説動画（YouTube）

  https://www.youtube.com/watch?v=uesNWFP40zw

その他見つけた比較情報

評価領域 / ベンチマーク	GPT-4o／GPT-4o mini	GPT-5／GPT-5 mini
MMLU（多領域言語理解）	GPT-4o：88.7% (ウィキペディア, Reuters)	明示的なスコア未公開。ただし GPT-5 はさらなる性能向上の有力候補（Passionfruit, TechRadar）
科学系 **GPQA Diamond** ベンチマーク	GPT-4o：70.1% (Vellum AI, Passionfruit)	GPT-5（Pro）：87.3%、GPT-5（coT ような reasoning モード）：高スコア（Passionfruit）
医療マルチモーダル推論（**VQA-RAD, SLAKE** など）	—（情報なし）	GPT-5：GPT-4o より大幅改善（最大 +20%）（arXiv）

MRI 脳腫瘍判断（画像＋推論）	GPT-4o：41.49% (arXiv)	GPT-5 mini：44.19%（最高）、GPT-5：43.71%、GPT-5 nano：35.85% (arXiv)
安全性・幻覚率	幻想率：1.49%（GPT-4o）、4o mini：1.69% (TechRadar)	GPT-5：1.4%（やや改善）(TechRadar)
コーディング能力（SWE-bench, Aider Polyglot）	データなし	GPT-5（thinking モード）：SWE-bench Verified 74.9%、Aider Polyglot 88% (Vellum AI, Passionfruit)
カスタムベンチマーク（labelstudio）	GPT-4o：0.88 精度 (Label Studio)	GPT-5 シリーズと「同等」精度を保持（GPT-5 base を上回る） (Label Studio)