生成 AI はなぜ“もっともらしい誤答”を生むのか
― 意味理解と確率モデルが生むハルシネーションの構造 ―

近年、私たちは必要な情報を検索する前に、まず「AI に聴いてみよう」と考えるようになった。曖昧な質問であっても、生成 AI は即座に応答し、まるで利用者の意図を理解しているかのように“それらしい答え”を返してくる。

しかし、この“もっともらしい回答”は、常に正しいとは限らない。むしろ、正答に見える誤答、いわゆるハルシネーションが潜んでいる可能性がある。これは単なる利用者の誤解ではなく、生成 AI の構造的性質に起因する問題である。

本レポートでは、Copilot に実際に問いかけながら、生成 AI がなぜ“もっともらしい誤答”を生むのか、その技術的・構造的要因を分析し、LLM と RAG の仕組みを体系的に整理する。

１．なぜ生成 AI は“もっともらしい誤答”を生むのか

生成 AI、とりわけ大規模言語モデル（LLM）は、事実と異なる内容をあたかも正しいかのように提示する「ハルシネーション」を起こすことがある。Copilot を含む主要モデルは、質問に対して「知らない」と明確に答えるよりも、学習したパターンに基づいて“もっともらしい答え”を構築する傾向を持つ。

この“もっともらしさ”は、利用者にとって自然で読みやすい一方で、誤答が正答のように見えてしまう危険性を内包している。これは特定モデルの癖ではなく、LLM という技術の構造的宿命である

ここでは、この宿命がどのような仕組みによって生じるのかを明らかにし、主要モデル間でどのような挙動の違いが現れるのかを体系的に分析する。

２．LLM の構造的限界：ハルシネーションの根源

２.１　LLM は「知識の有無」を判定できない（メタ認知の欠如）

人間は「知っている／知らない／あいまい／調べる必要がある」といったメタ認知を持つ。しかし LLM は、膨大なテキストの統計的パターンを学習したモデルであり、

・自分が何を知っているか
・何を知らないか
・どこが不確実か

を判定する仕組みを持たない。

したがって、LLM には「これは知らない」という自己認識が構造的に存在しない。

このメタ認知の欠如が、誤答を“もっともらしく”生成してしまう第一の要因である。

2.2　LLM は「意味」ではなく「確率」で文章を生成する

LLM は次の語を「意味」ではなく「確率」で予測するモデルである。

[例]「2026 年の中国の GDP は？」と問われた場合、LLM 内部では

・“2026 年”＋“中国”＋“GDP”という語の共起

・過去の学習データにおける GDP の典型的な数値表現

・文脈的に自然な語の並び

をもとに、最も尤もらしい数値を生成する。

このとき、「知らない」と答えるより、“それっぽい答え”の方が確率的に自然な文章になる。

これがハルシネーションの第二の要因である。

2.3　LLM は「知識の鮮度」を判定できない

Copilot の内部知識は 2025 年 8 月までで固定されている。しかし LLM はこの「知識の期限」を認識する仕組みを持たない。

そのため、「2026 年の出来事を問われても、古い知識を最新として回答してしまう」という構造的な誤りが発生する。

これがハルシネーションの第三の要因である。

2.4　RAG（検索）は補完するが、“もっともらしさ”は残る

Copilot は検索を積極的に使うため、最新情報の補完能力は高い。しかし、検索結果が曖昧な場合、LLM はその曖昧さを「もっともらしく補完」してしまうという性質を完全には消せない。

RAG はハルシネーションを軽減するが、LLM の構造的性質そのものを変えるわけではないため、根本的な解決には至らない。

3　Copilot の段階的補完メカニズム

Copilot は、LLM の構造的限界を補うために「内部知識 → 推論 → 検索 → 外部データ」という段階的な補完プロセスを採用している。検索は常に行われるわけではなく、必要と判断した場合にのみ発動する。この段階性が、Copilot の“最新情報対応力”を支えている。

3.1　補完プロセス

Copilot は、以下の順序で回答の確度を高めていく。

(1) 内部知識による判定

内部知識だけで高い自信度が得られる場合、検索は行わず内部知識で回答する。

(2) 推論エージェントによる補完

内部知識の組合せや推論によって回答可能と判断される場合、検索は不要となる。

(3) 検索（web_search）の発動

以下の条件のいずれかに該当すると、検索が発動する。

・内部知識の自信度が低い
・最新情報が必要
・固有名詞 / 日付 /
数値が曖昧
・利用者が「最新」「2026 年」などを明示
・内部知識と推論が矛盾
・情報が不足している

(4) 外部データソース

検索結果に加え、API・ニュース・統計などの外部データを参照し、回答の精度を高める。

3.2　検索の制約

Copilot の検索は万能ではなく、以下の制約のもとで運用される。

・検索結果の品質に依存する
・検索量は必要最小限に抑えられる
・未公表情報は取得できない
・信頼性が低い検索結果は破棄される

3.3　補完精度の判断基準

Copilot は、以下の基準をもとに回答の精度を判断する。

・内部知識の自信度（確率分布の集中度）
・検索結果の一致度
・ソースの信頼性
・情報の矛盾の有無

これらのプロセスと判断基準を総合的に評価し、最も信頼できる回答を構築する。

4　主要モデルの比較分析（2026 年版、Copilot 調べ）

主要な生成 AI モデルの挙動を、LLM の構造的性質と RAG（検索）統合の観点から比較する。比較軸は以下のとおりである。

・内部知識の自信度：内部知識をどれだけ優先するか
・検索の積極性：曖昧な質問に対して検索を発動する頻度
・最新情報の強さ：検索結果や外部データをどれだけ重視するか
・ハルシネーション傾向：不確実なときに“もっともらしい補完”を行う頻度
・最新 LLM（2026）：採用しているモデルの世代
・RAG の特徴：検索統合の仕組みと挙動

4.1　比較表

モデル	内部知識の自信度	検索の積極性	最新情報の強さ	ハルシネーション傾向	最新 LLM（2026）	RAG の特徴
Copilot（Microsoft）	中	高	高	中（検索で軽減）	2025 年夏モデル（内部知識は 2025 年 8 月まで）	多段階検索。最新情報対応力が非常に強い
ChatGPT（OpenAI）	高	中	中	高（内部知識優先）	GPT-4.1 / GPT-4o（2025）	検索は補助的。内部知識を強く信頼
Gemini（Google）	中	非常に高	非常に高	中〜高（検索ノイズ）	Gemini 2.0 Ultra（2025〜2026）	Google 検索と完全統合。最新情報最強
Claude（Anthropic）	低	低	低〜中	低（不確実なら答えない）	Claude 3.5（2025）	安全性重視。検索は控えめで誤答が少ない

4.2　モデル別の特徴

① Copilot（Microsoft）

・最新情報対応力が最も高い
・多段階検索により、内部知識の誤判定を検索で補正

・RAG 統合が実用的で、曖昧な質問にも強い

② ChatGPT（OpenAI）

・内部知識の自信度が非常に高い

・古い知識を最新として答えやすい

・ハルシネーションが最も起きやすい（内部知識優先のため）

③ Gemini（Google）

・検索統合が最強

・最新情報対応力は 4 社で最も高い

・検索ノイズを拾うと誤答が出る

④ Claude

・安全性重視で、不確実なら答えない

・ハルシネーションは最も少ない

・最新情報には弱い（検索を控えるため）

5　プロンプト設計の指針

生成 AI は、内部知識・推論・検索の組合せによって回答を構築するため、利用者がどの情報源を優先させたいかを明示することが、誤答を避けるうえで重要である。本章では、目的に応じてプロンプトをどのように調整すべきかを示す。

5.1　最新情報が必要なとき

最新の出来事・数値・統計・ニュースなど、内部知識だけでは不十分な情報を扱う場合は、検索を明示的に要求する。

[例]　「最新情報を検索して、内部知識と区別して説明して。」

[意図]　Copilot に検索を強制し、内部知識と外部情報を混同させない。
“もっともらしい古い情報”を排除する。

5.2　内部知識だけで体系的に説明させたいとき

歴史・概念・体系的知識など、検索ノイズを避けたい場合は、検索を使わないよう明示する。

[例]　「検索を使わず、内部知識だけで体系的に説明して。」

[意図]　LLM の内部表現（意味構造）を最大限活用
外部情報のノイズを排除
一貫した体系的説明を得る

5.3　内部知識と検索結果の両方が必要なとき

概念説明と最新情報の両方が必要な場合は、情報源の分離を明示する。

［例］　「検索結果と内部知識を分けて、それぞれの根拠を明示して説明して。」

［意図］　内部知識（体系）と検索結果（最新情報）を混同させない
情報源の透明性を確保
誤答の原因を特定しやすくする

6　生成 AI の“もっともらしさ”は技術の宿命である

生成 AI が“もっともらしい誤答”を生成するのは、Copilot 固有の癖ではなく、LLM という技術の構造的宿命である。以下の要因が複合的に作用し、ハルシネーションが生じる。

(1) メタ認知の欠如
自分が「何を知らないか」を判定できない構造

(2) 確率モデルによる文章生成
意味ではなく、統計的に自然な語の並びを優先する性質

(3) 知識の鮮度を判定できない構造
古い内部知識を最新情報として扱ってしまう

(4) RAG による補完の限界
検索で最新情報を補完しても、曖昧な結果は“もっともらしく補完”される

これらは LLM の内部構造に起因するため、完全に排除することはできない。

Copilot はこの宿命を上手く扱うモデルであり、検索を積極的に活用することで最新情報対応力を高めている。しかし、LLM の本質的性質は残り、ハルシネーションを完全に解消することはできない。

7　生成 AI は「理解」と「確率的生成」の二重性をもつ

生成 AI は、私たちが与えた問いを「理解している」ように振る舞う。しかし、その回答生成の仕組みは、常に確率モデルとしての性質を伴っている。LLM は内部表現として文脈や論理構造を保持し、意味的整合性を保ちながら応答を構築する一方で、最終的な出力は「次の語をどれだけ自然に続けられるか」という確率計算に基づいて生成される。

このため、生成 AI は次の二つの性質を同時に持つ。

・意味構造を保持し、文脈を理解する性質
内部表現として文脈・論理・概念の関係を捉え、整合的な説明を構築する。

・確率分布に基づき、もっともらしい語を選ぶ性質
統計的に自然な語の並びを優先し、不確実な部分を“それらしく補完”してしまう。

この二重性こそが、生成 AI が、ときに正確な説明を行い、ときに“もっともらしい誤答”を生む理由である。二つの性質は矛盾ではなく、LLM という技術の宿命として共存している。

生成 AI を正しく使いこなし、誤答を見抜き、教育・研究・社会実装において適切に扱うためには、この二重性を理解することが不可欠である。

以上