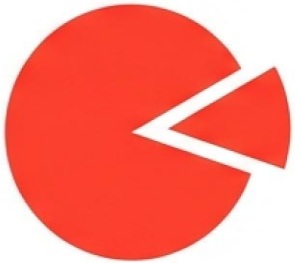
【第３回】 AIの「嘘」と「確率」の正体

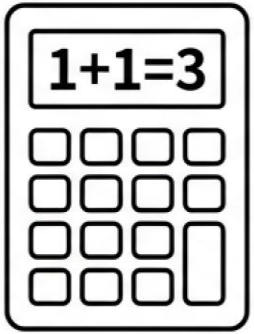
AIはなぜ、嘘をつくのか？

～ハルシネーションのメカニズムと、2つのエンジニアリングによる対策～



【現象】こんな「もっともらしい嘘」つかれたことありませんか？

存在しないURLや、
架空の論文を提示された



計算間違いを
自信満々に回答された

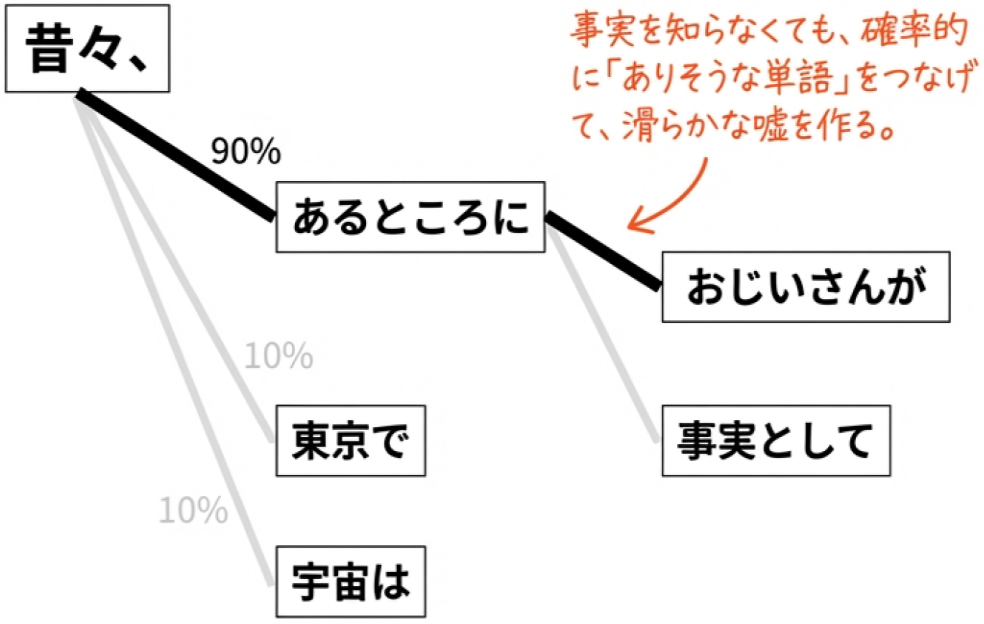
マニュアルにない手順を
勝手に捏造された

これらはバグではありません。生成AIの仕様です。
この現象を専門用語で「ハルシネーション（Hallucination：幻覚）」と呼びます。

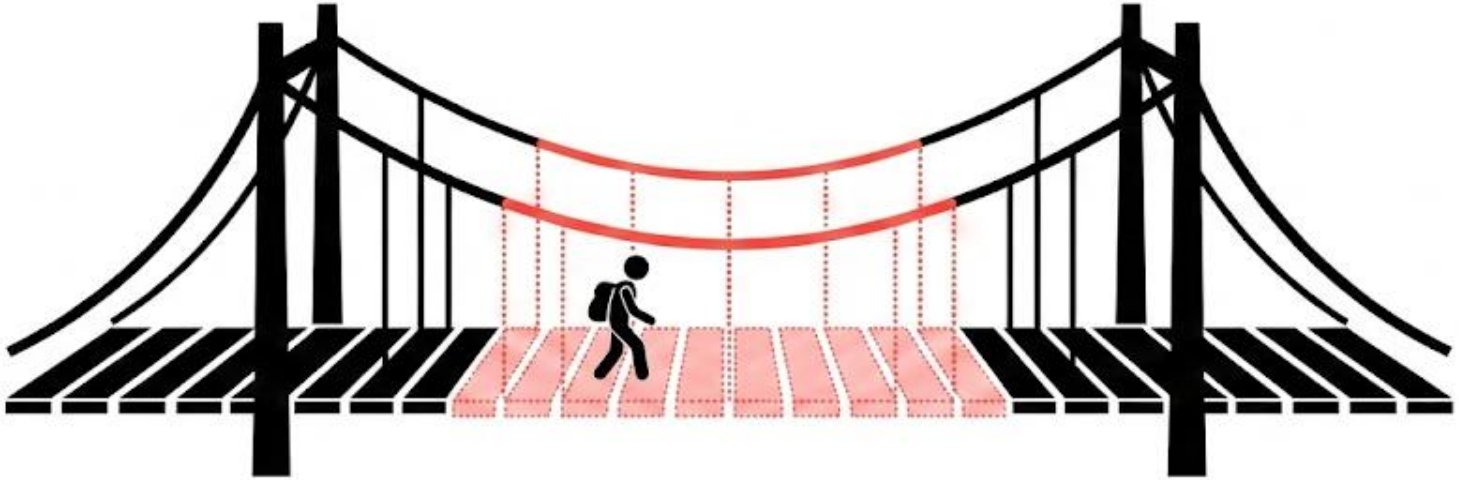
【原因①】AIは「事実」を知らない（確率的生成）

第1回の復習：
AIの正体は「巨大な辞書」ではなく、「次に来る確率が高い言葉を予測してつなげるマシン（Next Token Prediction）」です。

嘘のメカニズム：
AIにとっての「正解」とは、「事実かどうか」ではなく、「文章として自然に繋がっているか（もっともらしいか）」です。



【原因②】「知識の空白」を埋めようとする親切心



「分かりません」と言わない、行き過ぎたサービス精神

知らない事を聞かれた際回答を出さないよりも、不正確でも答えを出す方が「良い回答」だと学習しています。そのため、「知りません」とは言えず、情報を捏造してでも回答を出します。

【分類】嘘には「2つの種類」がある

	Type A:「知識不足」による捏造	Type B:「読み間違い」による誤認
状態 (State)	0→1の捏造 (情報がないのに嘘をつく)	参照ミス (資料はあるのに間違える)
原因 (Cause)	学習データに知識がない	記憶（コンテキスト）が汚れている 読み飛ばしている
対策 (Fix)	情報を与える (Context Engineering)	読み方を縛る (Prompt Engineering)

【対策①】グラウンディング（情報の提供）
～これは「コンテキストエンジニアリング」の実践です～

※第二回内容実践

Before

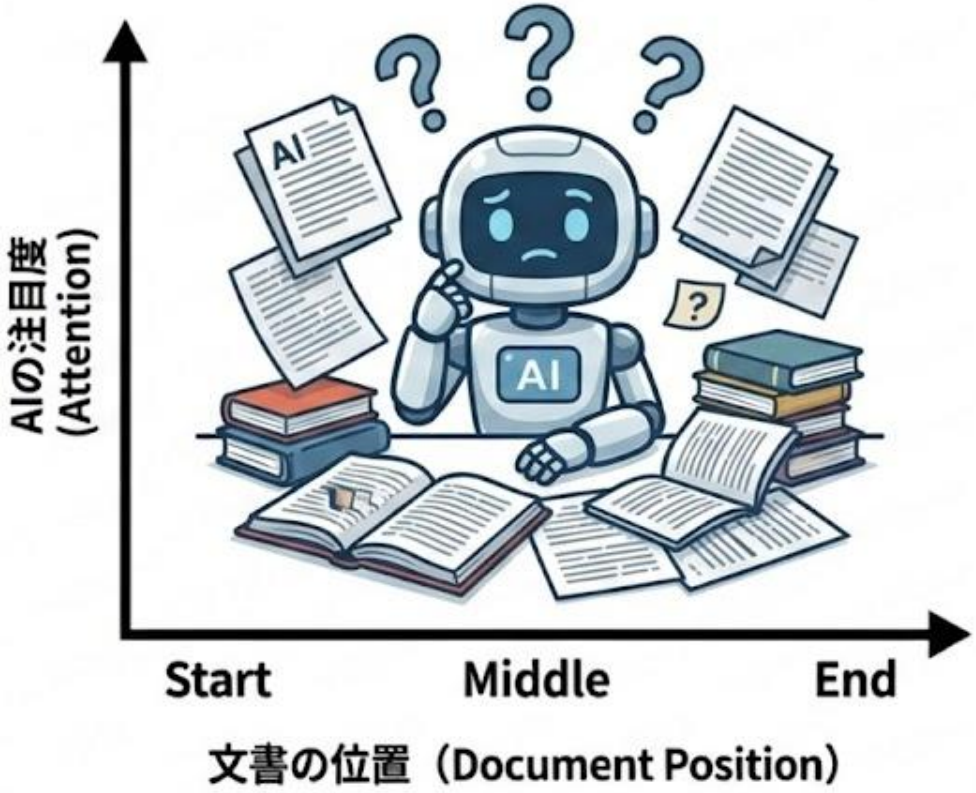
学習データのみ（記憶頼り）

After

ファイル添付（社内資料、PDF、議事録データ）

Grounding: AIのあやふやな脳みそを使わせず、
「手元の資料だけを見て答えろ」と指示する技術です。

【注意点】「ただ渡すだけ」では、AIは読み飛ばす



情報過多: コンテキストウィンドウが広くても、AIは全てを均等に読めません。

解決策: 不要なノイズを削除し、「本当に必要な情報」だけを選別して渡すことが重要です。

【対策②】「推測禁止」と「不知の表明」

～これは「プロンプトエンジニアリング」の実践です～

※第一回内容実践

プロンプト例

```
# 制約事項（重要）
- 回答は必ず添付ファイルの記述のみに基づきなさい。
- ファイルに記載がない情報は、自分の知識や想像で補完してはいけません。
- 情報が見当たらない場合は、無理に答えず、一言「記載なし」とだけ回答してください。
- 絶対に推測で情報を補完しないでください。
```

制約(Constraint): 資料を渡すだけでは、AIは想像で補おうとします。この親切心をルールで封じ込めます。

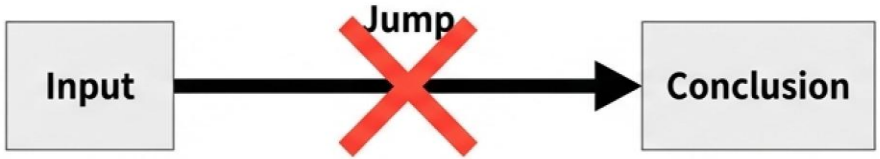
【対策③】出典（Source）の明示と「検証可能性」

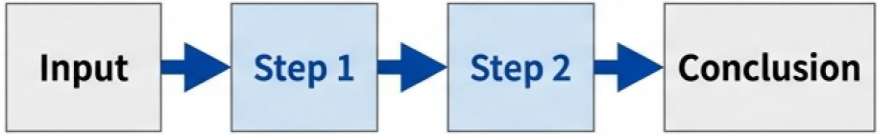
プロンプト例

```
# 制約事項（重要）
-**回答の（各項目）には、必ず根拠となった「ファイル名」と「ページ数」を付記してください。**
- ユーザーに「どこを見てそう言ったのか？」が明確に伝わるように回答を構成してください。
```

Transparency: これは「AIにミスをさせない」技術ではなく、
「人間がAIの嘘を見抜く」ためのリスク管理です。
「どこを見てそう言ったのか?」を提示させることで、ファクトチェックが可能になります。

【対策④】
思考の過程を書かせる（Chain of Thought）

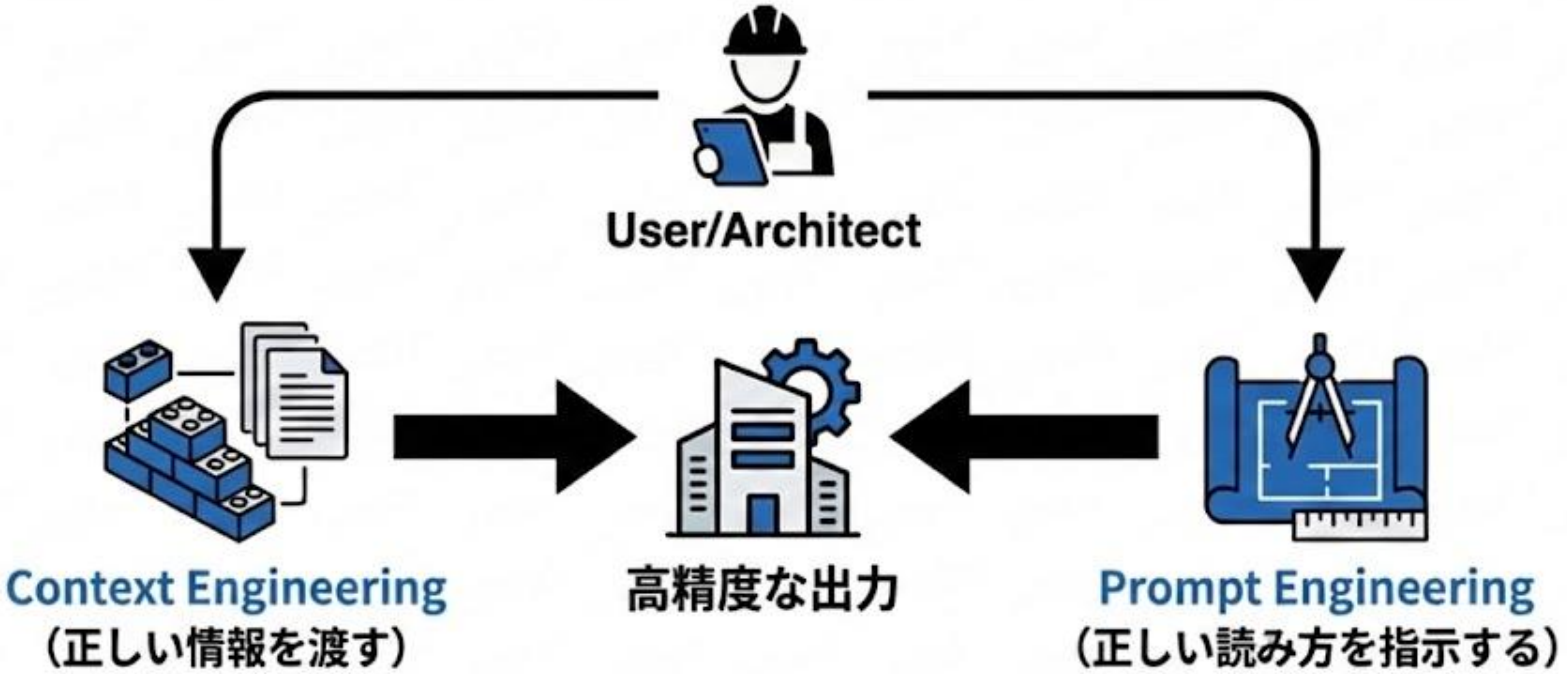




結論を出す前に、参照した情報の分析プロセスをステップごとに記述してください。

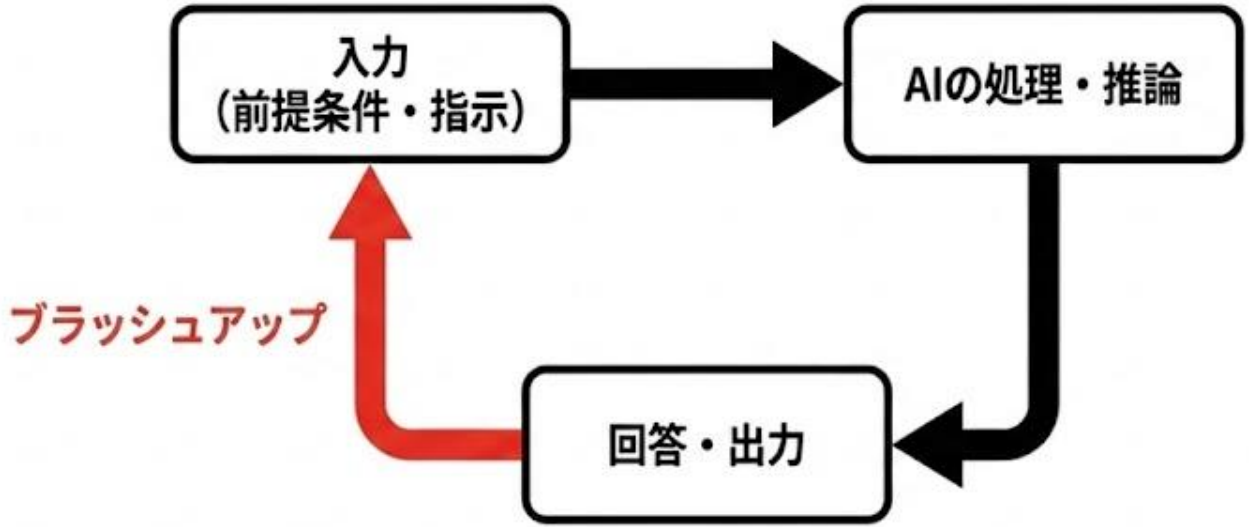
Step-by-Step: いきなり結論を出させると、AIは直感（確率）で答えがちです。推論プロセスを出力させることで、論理の飛躍を防ぎます。

【総集編】信頼性は「使い手の設計スキル」で決まる



AIの嘘を防ぐ特効薬はありません。
様々な技術を重ね合わせて信頼性を高めることが重要です。

まとめ：AIに疑いをかける前に、「設計」を見直す



Reframing: ハルシネーション（嘘）はAIの特性であり、100%正しいとは限りません。
しかし、情報や指示を具体化することで、そのリスクを抑え強力なパートナーとなります。

Action: 「嘘ばかりで使えない」と切り捨てる前に、プロンプトやコンテキストの設計を見直してみましょう！AIへのアプローチを最適化することが、業務効率化への近道です。

Next Step：手動から自動へ。AIに「外部脳」を接続する

信頼性（中身）は確保できました。しかし、毎回手動でファイルを渡すのは非効率ですし、手間です。
次回は、AIが必要な情報を「自ら検索し、読み込み、回答する」システムについて学んでいきましょう。

【第4回】検索拡張生成（RAG）の基礎 ～自社データを「カンニング」させる技術～