Structural Factors of AI Hallucinations

Six Perspectives for Essential Understanding

AI Utilization Guide for Small and Medium Enterprise Consultants

Certified SME Management Consultant (Japan), member of the Jonan Branch.Member of the Generative AI Research Group (Jonan Branch).
Member of the Small M&A Research Association (General Incorporated Association).

Why Understanding AI Structure is Essential

In the work of small and medium enterprise consultants, the use of AI is rapidly increasing across financial analysis, market research, and business planning. However, many users hold the misconception that "AI is an advanced knowledge database."

In reality, AI is a **probability-optimized text generator**. Using AI without understanding this essential nature means making decisions based on structurally incomplete information.

Foundational Recognition: Hallucinations Are Not "Bugs"

Common Misconception

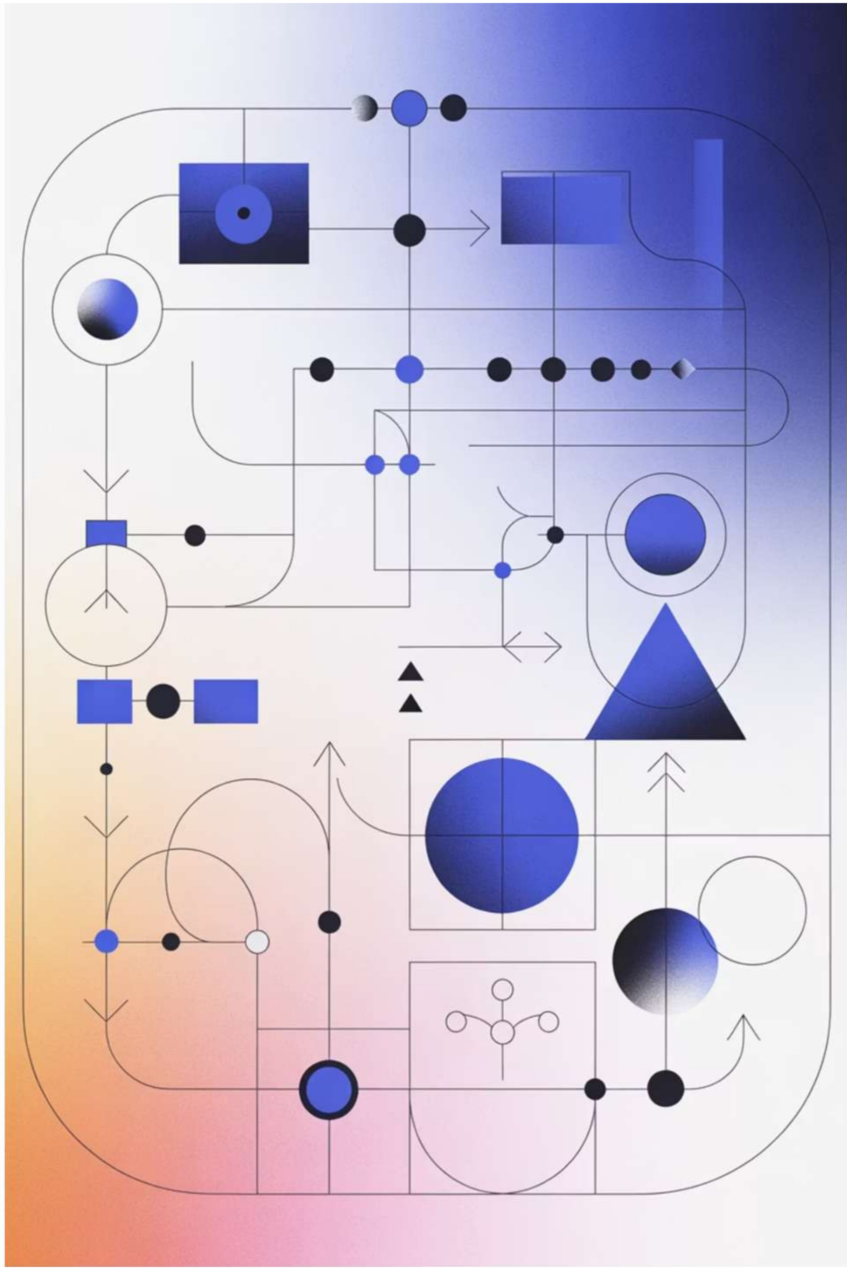
Hallucinations are defects or malfunctions in AI systems that need to be fixed through better engineering.

The Reality

AI is operating exactly as designed according to its specifications and training objectives.

The core issue is the **structural mismatch between "user expectations" and "model probability optimization."** Users expect "accurate information based on facts," while the model performs "generation of text that is most probabilistically natural given the context."

This fundamental disconnect creates inevitable challenges that cannot be eliminated through technological refinement alone—they require structural understanding and appropriate workflows.



Perspective 2: Reconstruction from Internal Compressed Representations

Traditional Databases

- Reference fixed storage locations
- Static retrieval process
- Return only recorded information
- Exact reproduction guaranteed

LLM Vector Space

- Dynamically decompose and reconstruct
- Flexible processing mechanisms
- Probabilistically generate new combinations
- Creative synthesis possible

LLMs store training data in "**compressed vector space**" representations. Words with similar meanings are positioned close together in this multidimensional space. However, this proximity can lead to **concept confusion** when vectors are too close, and **incorrect pattern chains** when probabilistic associations are formed inappropriately.

Unlike databases that retrieve exact records, LLMs reconstruct responses from compressed semantic representations, introducing opportunities for drift and hallucination during the reconstruction process.

Perspective 3: Separation of Confidence and Accuracy [Structural Deception]

Human Communication

When uncertain, people hedge their language and use qualifiers. When confident, they speak definitively. Incorrect answers correlate with "knowledge uncertainty."

LLM Behavior

LLMs select the most probabilistically natural expression based on context, independent of information accuracy. Confidence in tone does not reflect accuracy of content.

The Reversal Phenomenon

Incorrect information may be generated with complete confidence, while **accurate information may be expressed tentatively**. This creates "structural deception" that contradicts human intuition, making it critical to separate tone from factual verification.

This disconnect between presentation confidence and content accuracy represents one of the most dangerous aspects of LLM outputs for professional use. Users must develop systematic verification processes that ignore confidence signals.

Perspective 4: Absence of "World Models" and Weak Causal Reasoning

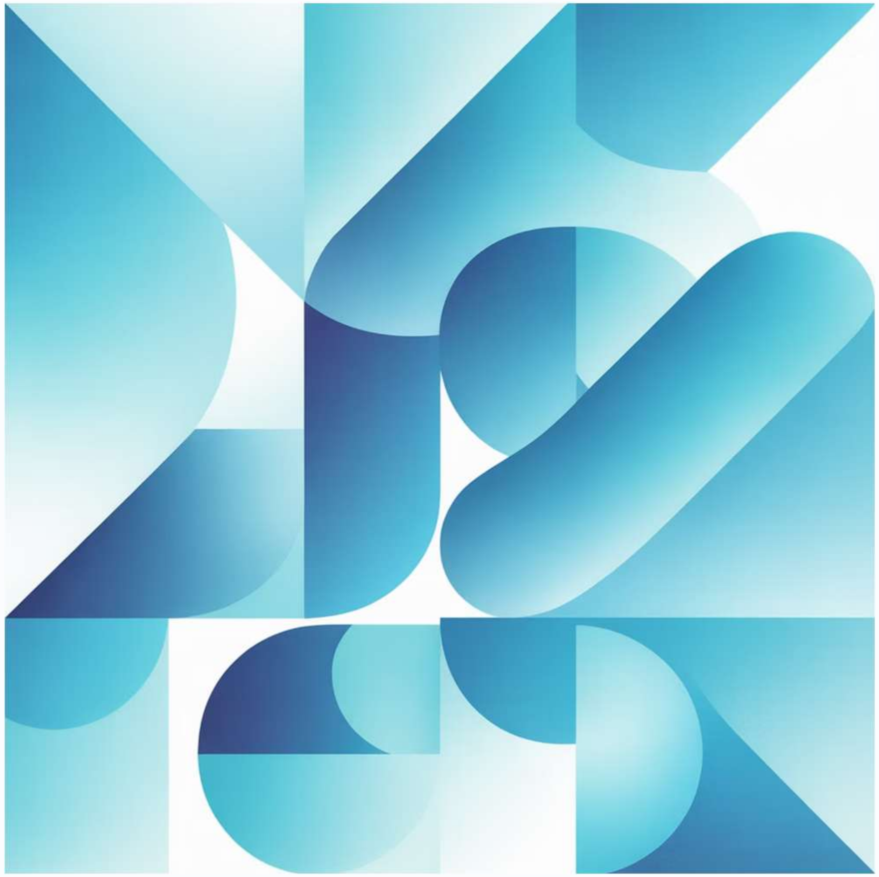
LLMs do not understand causal relationships, time sequences, or physical laws as "meanings"—they output them as **statistical patterns in text**.

Human Understanding

"Ice melts when warmed" → Understood as a physical law based on thermodynamic principles and molecular behavior

LLM Processing

Learns co-occurrence patterns of words "ice," "warm," and "melt" from training data and probabilistically reproduces them without understanding the mechanism



Advantages

Disadvantages

Perspective 5: The Impact of Ambiguity

The hallucination occurrence rate strongly depends on **the ambiguity (not abstraction level) of user instructions**. Higher ambiguity expands the inference range, dramatically increasing the probability of incorporating fictional elements (noise) into responses.

Low Ambiguity Instruction

"What is the travel time from Tokyo to Osaka by Shinkansen?"

→ Specific factual information with narrow inference range and high accuracy potential

Medium Ambiguity Instruction

"What are methods to reduce Shinkansen travel time?"

→ Several possible answers, wider inference range with moderate risk

High Ambiguity Instruction

"What is the optimization of Japan's transportation system?"

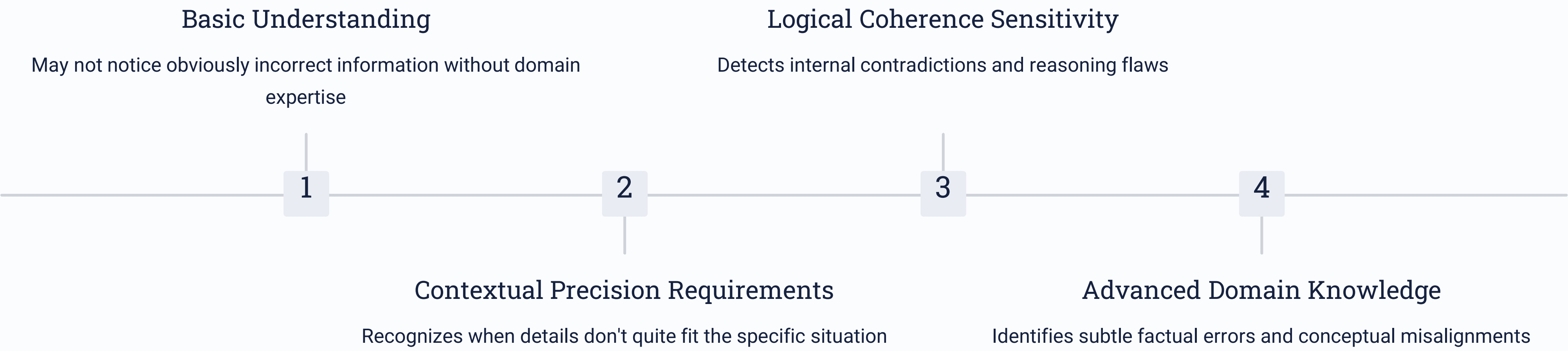
→ Very wide inference range where fiction easily contaminates the response

To minimize hallucination risk, always provide specific context and narrow the scope of inquiry. Frame questions with clear boundaries, concrete parameters, and verifiable criteria whenever possible.

Perspective 6: Relationship Between User Understanding and Hallucination Recognition

Visualization Capability

"Visualizing" hallucinations is not possible for everyone. The higher the user's expertise level, the more visible subtle errors become. Detection capability is fundamentally limited by the user's own knowledge domain.



1 Basic Understanding

May not notice obviously incorrect information without domain expertise

2 Contextual Precision Requirements

Recognizes when details don't quite fit the specific situation

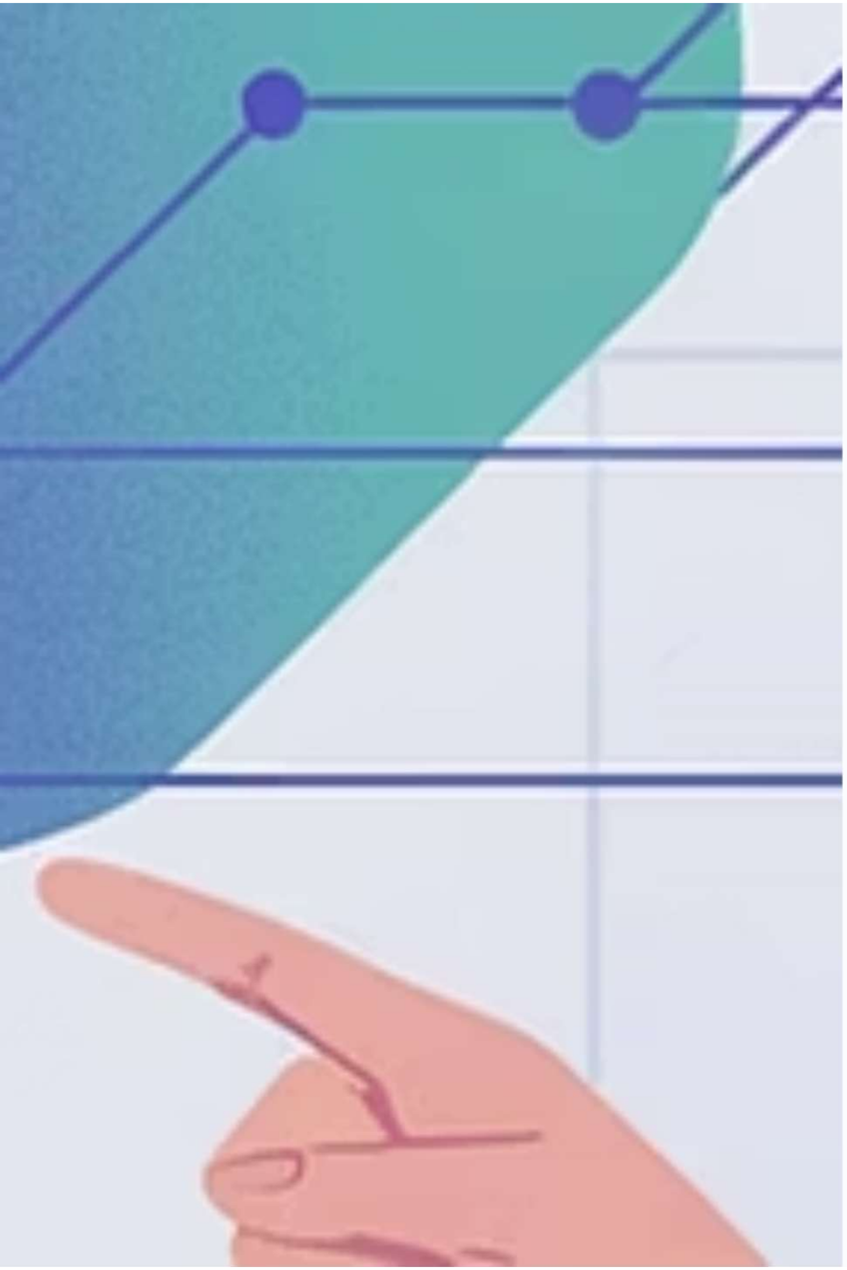
3 Logical Coherence Sensitivity

Detects internal contradictions and reasoning flaws

4 Advanced Domain Knowledge

Identifies subtle factual errors and conceptual misalignments

Users with shallow domain expertise may not detect even obviously incorrect information. Hallucination detection ability **depends on the user's own understanding and literacy**. This creates a dangerous situation where those who most need verification assistance are least equipped to perform it.



Perspective Transformation Through Structural Understanding

✗ Incorrect Recognition

- "AI is a knowledge database"
- "Better accuracy will solve the problem"
- "Hallucinations are technical defects"
- "AI can be trusted like an expert"

○ Correct Recognition

- "AI is a probability-optimized text generator"
- "Verification processes are structurally mandatory"
- "Hallucinations are consequences of design specifications"
- "AI is a tool requiring expert oversight"

By deepening understanding through these six perspectives, your fundamental approach to AI changes. With this structural understanding as a foundation, appropriately leveraging AI becomes a professional responsibility for consultants and business advisors.

This knowledge transforms AI from a mysterious "black box" into a well-understood tool with known characteristics, limitations, and appropriate use cases. Professional practice demands this level of comprehension.

Five Usage Patterns That Amplify Hallucinations and Countermeasures

When using LLMs in professional practice, highly abstract instructions tend to trigger hallucinations (generation of non-factual information). This section provides detailed explanations of five common dangerous patterns, their structural causes, and practical countermeasures.

Understanding these patterns is essential for safe and effective AI utilization. Each pattern represents a structural vulnerability in how LLMs process information, and awareness allows you to design workflows that minimize risk while maximizing value.

Pattern 1: Continuing Deep Causal Investigation

What Happens

Repeatedly asking "Why?" "What's the reason?" "What's the deeper background?" causes hallucinations to surge dramatically.

Structural Reason

LLMs have limitations in causal reasoning. The deeper you dig, the more they must probabilistically generate "causal hierarchies" that don't exist in training data.

Question 1: Sales Decline Reason

"Market environment changes"

Safety: High

Question 2: Market Change Reason

"Diversification of consumer needs"

Safety: Medium

Question 3: Diversification Reason

"Due to SNS proliferation..."

Safety: Low

Question 4: Further Investigation

Completely fictional causal chain

Safety: Dangerous

Countermeasure

Limit causal investigation to 2-3 levels. Beyond that, start a new thread and verify the generated causal relationships at each stage before proceeding to the next level of inquiry.

Pattern 2: Abstraction to the Extreme Limit

Asking for extreme abstraction such as "What is the essence?" "What is the universal principle?" "What is the generalization?" causes LLMs to become unstable. This occurs because such requests exceed the stable regions of the vector space.

Stable Domain Questions

"What are the challenges in Project A?"

→ Has specific context, allowing relatively safe responses with grounded information

Unstable Domain Questions

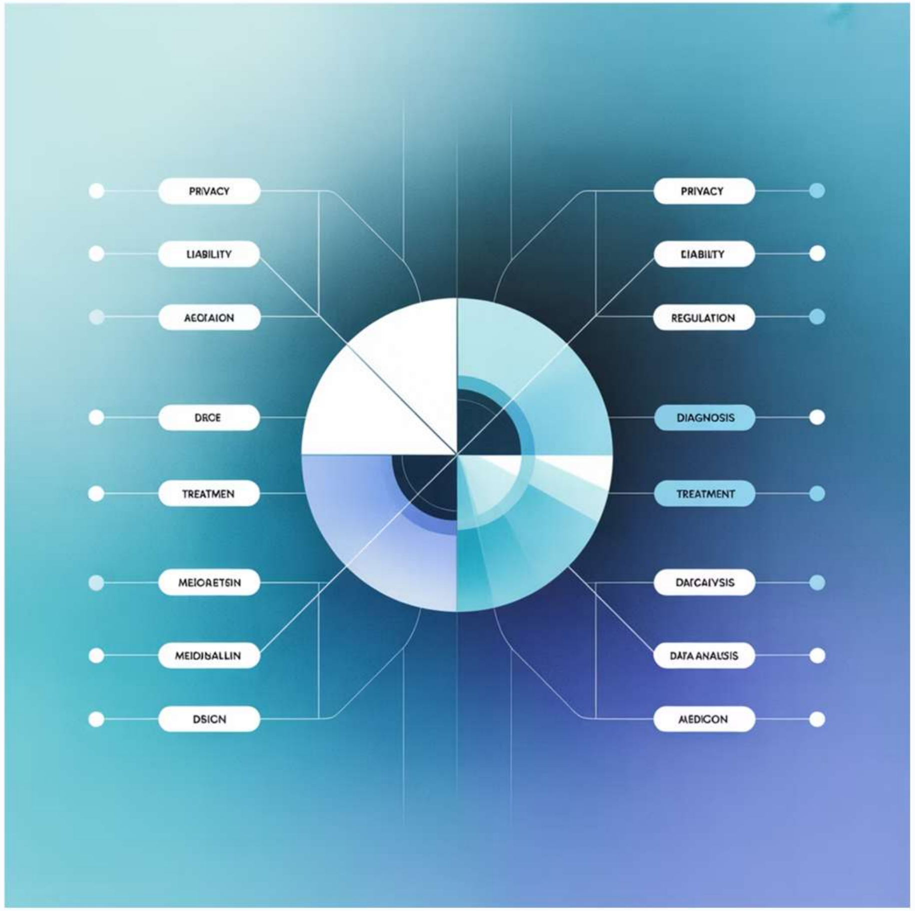
"What are the essential challenges common to all projects?"

→ Searches broad areas of vector space, making fictional content contamination highly likely

Why This Is Dangerous

LLM internal representations are most stable when tied to specific contexts. As abstraction level increases, the system must draw from broader, less certain areas of vector space, reducing accuracy proportionally.

Pattern 3: Integrating Multiple Expert Domains



Requesting integration across multiple expert domains such as "analyze from medical × AI × legal perspectives" causes hallucinations to surge. This triggers **semantic interference** in the concept space.

Different expert domains are arranged in "different vector regions" within the LLM's internal structure. Forcing their integration causes unrelated concepts to be probabilistically connected, generating fictional relationships.

Single Domain Question

"What are the legal implications of AI in healthcare?"

Focuses on one domain with clear boundaries and established patterns

Multi-Domain Integration Request

"Synthesize medical ethics, AI capabilities, and legal frameworks into unified recommendations"

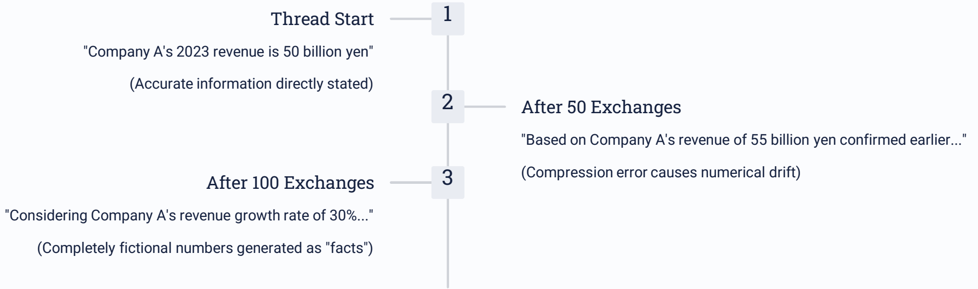
Forces artificial connections between separate knowledge domains, creating unstable outputs

Countermeasure

Analyze each expert domain separately, then perform integration through human judgment. Use different threads for different domains to maintain conceptual boundaries and prevent semantic collision.

Pattern 4: Continuous Use in Long Conversation Threads

Continuing a large volume of exchanges in a single thread (conversation) for an extended period causes hallucinations to gradually increase. This results from **context compression loss**.



1 Thread Start

"Company A's 2023 revenue is 50 billion yen"

(Accurate information directly stated)

2 After 50 Exchanges

"Based on Company A's revenue of 55 billion yen confirmed earlier..."

(Compression error causes numerical drift)

3 After 100 Exchanges

"Considering Company A's revenue growth rate of 30%..."

(Completely fictional numbers generated as "facts")

Why This Occurs

While LLMs retain entire conversation history, as conversations lengthen, early context information becomes compressed and precision decreases. Furthermore, the LLM begins referencing its own previously generated (unverified) content as "fact" in a compounding error cycle.

Pattern 5: High-Density, High-Speed Continuous Use

What Happens

Throwing a large volume of questions in rapid succession causes response quality to deteriorate rapidly. This phenomenon is called **premature convergence** or "exploration shortcut."

Slow Usage Pattern

Question 1 → Response (deep exploration)

(5-minute wait)

Question 2 → Response (deep exploration)

Each response benefits from thorough exploration

High-Speed Continuous Pattern

Question 1 → Response

Question 2 → Response (shallow, extends Q1)

Question 3 → Response (converges to pattern)

Question 4 → Response (nearly mechanical)

Originality decreases, accuracy deteriorates

Countermeasure

Allow temporal spacing between questions to maintain LLM exploration quality. Avoid rapid-fire continuous use. When multiple related questions are needed, batch them into a single well-structured prompt rather than sequential rapid queries.

Comparison and Risk Levels of Five Patterns

Pattern	Risk Level	Detection Difficulty	Primary Countermeasure
Deep Causal Investigation	Very High	High	Limit to 2-3 levels, verify each stage
Extreme Abstraction	Very High	Very High	Maintain specificity, use concrete examples
Multi-Domain Integration	High	High	Analyze domains separately, integrate manually
Long Thread Usage	Medium	Medium	Reset threads every 20-30 exchanges
High-Speed Continuous	Medium	Low	Space out questions, allow processing time

Each pattern presents distinct challenges requiring different mitigation strategies. The highest-risk patterns (deep causal investigation and extreme abstraction) also pose the greatest detection challenges, making them particularly dangerous for users with limited domain expertise.

Comprehensive Overview of Practical Countermeasures

Staged Approach

Limit causal investigation and abstraction to 2-3 stages. Beyond that, restart with a new thread and verify at each stage before proceeding.

Maintain Specificity

Avoid abstract questions and always provide specific context. Limit scope as in "regarding △△ in the situation of ○○."

Domain Separation

Analyze multiple expert domains individually, with integration performed by human experts. Use different threads for different domains.

Regular Reset

Start new threads after approximately 20-30 exchanges. Always explicitly restate important numbers and facts in each new context.

Appropriate Spacing

Allow temporal intervals between questions to maintain LLM exploration quality. Avoid high-speed continuous use patterns.

Continuous Verification

Always verify generated responses. When using dangerous patterns, exercise particular caution and implement systematic fact-checking.

Practical Application Checklist

Pre-Use Verification Items

- **Question Specificity:** Is the question too ambiguous?
- **Causal Depth:** How many levels of causation are being requested?
- **Number of Expert Domains:** Are you attempting to integrate multiple specialized fields?
- **Thread Length:** How many exchanges have occurred in the current thread?

Post-Use Verification Items

- **Fact Verification:** Are generated numbers and facts accurate?
- **Causal Validity:** Are presented causal relationships logical and well-supported?
- **Expert Knowledge Confirmation:** Has specialized content been clearly verified?
- **Consistency Check:** Are there contradictions between responses?

By utilizing these checklists in practical work, you can significantly reduce hallucination risks. For particularly important projects, make these verification items mandatory steps in your workflow.

Consider creating standardized verification protocols for your organization that incorporate these elements, ensuring consistent quality across all AI-assisted work products.

Summary: Principles for Safe LLM Utilization

01

Recognize the Five Dangerous Patterns

Deep causal investigation, extreme abstraction, multi-domain integration, long thread usage, high-speed continuous use

02

Understand Structural Reasons

Grasp the mechanisms of why each pattern triggers hallucinations and how the underlying architecture creates these vulnerabilities

03

Implement Appropriate Countermeasures

Staged approach, maintaining specificity, domain separation, regular resets, appropriate spacing between queries

04

Continuously Verify

Always critically evaluate generated content and obtain expert confirmation when necessary, especially for high-stakes decisions

LLMs are powerful tools, but understanding their limitations and using them appropriately is crucial. By avoiding these five patterns and practicing recommended countermeasures, you can minimize hallucination risks and leverage LLMs safely and effectively in professional practice.

When using LLMs in practical work, sharing these guidelines across your organization and ensuring all team members understand safe usage methods is the key to success. Establish organizational standards and verification protocols that embed these principles into standard operating procedures.