# AI Guardrails: Spotting & Fixing Common LLM Failure Modes

A Free Practical Guide from Libra Sentinel's *AI Literacy Series*

# Common Ways AI Models Slip Up (and How to Spot Them)

★ Large Language Models (like ChatGPT) are powerful, but they also have predictable blind spots.

★ These patterns of mistakes (also called 'Failure Modes'), don't mean the AI is broken, they're just part of how the technology works.

★ Knowing what they look like helps you catch errors quickly and use AI more effectively.

| Failure Mode | What It Is | When It Happens | Effect on Output | User Signals | Prevention / Mitigation |
|---|---|---|---|---|---|
| **Hallucination/ Overconfidence ✨** | Fabrication: model invents facts, sources, or reasoning. | Can occur at any length; more likely when asked for specifics it doesn't know, or when context is ambiguous. | Plausible-sounding but false content, delivered with high confidence. | Outputs look polished but don't check out on verification. | Add disclaimers ("don't guess"); force citations; ask for uncertainty signals; verify with external sources. |
| **Lost in the Middle🌀** | Soft degradation: model pays less attention to mid-context information. | Even within the token limit, especially with 10k–50k tokens in long documents. | Confidently misses or distorts details buried in the middle. | Answers are accurate about intro/conclusion, vague/wrong on middle parts. | Use retrieval-based prompting; ask section-specific questions; highlight/anchor key details in the prompt. |
| **Truncation ✂️** | Hard cutoff: the model drops part of the input that exceeds the context window. | When total input (your text + instructions + expected output planning) > model's max tokens. | Entire sections (usually at the start) are ignored. | Output ignores early parts of your text; summary feels incomplete. | Keep inputs well below max limit;<br>chunk documents;<br>summarize in stages |

Part of the AI Literacy Series (Courses: Understanding LLMs | Prompt Fluency Toolkit | Ethical Prompt Fluency)

# AI Guardrails: HALLUCINATION + OVERCONFIDENCE

*How to reduce fabricated answers and rein in the model's tendency to sound more certain than it really is.*

## Why It Matters

LLMs can produce text that looks polished but is factually wrong. Worse, they often present these errors with high confidence. This is the classic "hallucination + overconfidence" problem. Guardrails help reduce risks in compliance, legal, and high-stakes contexts.

## Use Guardrails When Accuracy Matters For:

- Fact-heavy tasks (laws, compliance, contracts, medical).
- Any document shared externally.
- When prompts are vague or under-specified.
- Long/complex outputs at risk of distortion.

| Technique | Copy-paste prompts |
|---|---|
| **Add disclaimers / no guessing** | 1) *"Do not guess. If uncertain or missing context, reply '[UNCERTAIN]' and state what you need."*<br>2) *"Work strictly from [DOC]. If the answer isn't in [DOC], say '[NOT IN SOURCE]'."* |
| **Require citations / anchors** | 1) *"For every claim, cite [URL/title/section] or write 'No source'."*<br>2) *"Quote exact lines/paragraphs from [DOC] for each claim; no external info."* |
| **Ask for uncertainty signals** | 1) *"After the answer, add 'Confidence: X/10' and list 2 reasons."*<br>2) *"List your top 3 assumptions and mark each [LOW]/[MED]/[HIGH] confidence."* |
| **Cross-check / verification step** | 1) *"Give the answer, then suggest 2 reputable sources to verify it; note likely points of disagreement."*<br>2) *"Compare [SOURCE A] vs [SOURCE B] on [QUESTION]; summarize agreements, conflicts, and what remains uncertain."* |
| **Break into smaller, verifiable steps** | 1) *"Answer in three sections: Assumptions → Evidence from [DOC] → Conclusion."*<br>2) *"First outline 3–5 steps you'll take; pause for 'Go' before proceeding."* |

## When You Can Relax Guardrails

- Brainstorming, ideation, and creativity.
- Personal drafting (journals, note summaries).
- Low-stakes first drafts.
- Exploratory "what if" scenarios.
- When you're confident in spotting/reviewing errors.

## Key Takeaway

AI outputs can look polished but still be wrong. Guardrails don't eliminate hallucinations — they make errors easier to detect and correct before they reach critical use.

# AI Guardrails: LOST IN THE MIDDLE

*How to reduce errors when models miss or distort mid-context details.*

### Why It Matters

LLMs often give too much weight to the beginning and end of your input while ignoring what's buried in the middle — especially in long prompts (10k–50k tokens). This can lead to summaries or answers that sound right but skip or misrepresent key mid-context details. That's risky in compliance, contracts, or technical documents where "the middle" often carries the critical information.

### ❗ Use Guardrails When Accuracy Matters For:

- Long or complex documents (policies, contracts, reports).
- Step-by-step reasoning chains with crucial middle evidence.
- Research summaries where nuance is spread across sections.
- Any workflow where omissions could mislead.

| Technique | Copy-paste prompts |
|---|---|
| **Retrieval-based prompting** | "Answer using only sections 3–5 of this document.""What are the key arguments in the *middle* of the text (not intro or conclusion)?" |
| **Section-specific questions** | "Summarize each section separately: [Intro], [Methods], [Findings], [Conclusion]." |
| **Anchoring key details** | "Highlight the 3 most important facts from paragraphs 10–20." |
| **Force coverage** | "List one point from the beginning, one from the middle, and one from the end of this text." |
| **Step-wise synthesis** | "First summarize sections A, B, C separately. Then combine into a full summary." |

### ✅ When You Can Relax Guardrails

- Short documents (well under model context limits).
- Brainstorming or exploratory "big picture" summaries.
- Low-stakes uses where omissions aren't critical.

### Key Takeaway

LLMs can skip or distort mid-context content. Guardrails help by forcing coverage of middle sections and breaking work into smaller, section-anchored steps.

# AI Guardrails: TRUNCATION

*How to reduce errors when inputs exceed the model's context window.*

## Why It Matters

When your input (instructions + text + expected output planning) is larger than the model's maximum token limit, the model will silently drop part of the text, usually at the start. This leads to summaries or answers that ignore critical context, even if the output looks polished.

## ❗ Use Guardrails When Accuracy Matters For:

- Long policies, contracts, or multi-section reports.
- Legal/compliance reviews where every section matters.
- Any workflow where the intro/background contains essential definitions.
- Multi-step reasoning tasks across large inputs.

| Technique | Copy-paste prompts |
|---|---|
| **Chunking input** | "Summarize pages 1–3 first. Then we'll continue with 4–6." |
| **Stage summaries** | "Give a detailed summary of this section only. We'll combine summaries later." |
| **Prioritize scope** | "Focus only on [Topic A] in this document. Ignore other sections." |
| **Sliding-window review** | "Process this text in overlapping chunks of 1000 words; carry forward context each step." |
| **External retrieval** | "If the document is too long, ask me to provide a smaller section instead of skipping content." |

## ✓ When You Can Relax Guardrails

- Short documents or prompts well under the model's token limit.
- Creative/brainstorming tasks where exact coverage isn't needed.
- Quick high-level summaries where detail loss is acceptable.

## Key Takeaway

Truncation doesn't distort content, it drops it entirely. Guardrails help by chunking inputs, staging summaries, or forcing narrower scope so critical sections aren't lost.

This free guide is part of Libra Sentinel’s AI Literacy Series. Explore our courses on Understanding LLMs and Prompt Fluency Toolkit™  to build real-world AI skills.