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; chunk documents; summarize in stages

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