Understanding Why Generative AI Sometimes Gets Things Wrong

When a generative AI tool produces an answer that sounds confident but is completely wrong, the experience can be unsettling. The model might invent a historical event, misstate a simple fact, or cite a source that does not exist. This behavior, often called an AI hallucination, is one of the most important concepts to understand for anyone using these systems at work.

Knowing why these errors happen and how to catch them is essential for using generative AI safely. This article explains the root causes of these mistakes, offers practical ways to spot them, and gives you concrete steps to reduce the risk of spreading incorrect information in your organization.

What Causes a Model to Make Up Information?

Generative AI models are not databases or search engines. They are pattern-matching engines trained on enormous amounts of text. When you ask a question, the model predicts the most likely sequence of words based on everything it has seen during training. It does not have a stable internal record of facts. It has a statistical map of language.

This design means the model can produce a perfectly grammatical and plausible answer that contains no truth at all. The model is not lying. It simply does not know what it does not know. The training process encourages the model to produce coherent responses, not necessarily accurate ones. When the model lacks the specific pattern for a correct answer, it still generates something. That something can be an AI hallucination.

View the location on Google Maps

Another factor is that these models compress information. They learn general relationships, not exact copies of data. Details that appear infrequently in training can get blurred or lost. When a model guesses at a missing detail, it fills the gap with something that fits the pattern, even if that something is invented.

Common Types of Errors in Practice

Errors from generative AI fall into a few recognizable categories. Recognizing them helps you decide how much to trust the output.

- Factual mistakes: The model states a date, name, or number that is wrong. For example, it might say a company was founded in 2005 when it was actually founded in 2010.

- Source fabrication: The model cites a study, book, or article that does not exist. It may give an author name and a title that sound real but are entirely made up.
- Logical contradictions: The model contradicts itself within the same response. It might say a process takes three steps and then describe four steps later in the same paragraph.
- Context confusion: The model mixes up details from different but related topics. It might answer a question about one software feature with steps that belong to a completely different feature.

These categories overlap. A single response can contain multiple types of error. The key point is that none of these errors are obvious from the phrasing. The model delivers them with the same confident tone it uses for correct answers.

Why This Matters in the Workplace

When teams use generative AI for drafting emails, summarizing reports, writing code, or answering customer questions, a single hallucinated fact can have real consequences. A wrong date in a client proposal can damage credibility. A fabricated legal citation can create compliance risks. A confident but incorrect code snippet can introduce bugs that take hours to find.

The risk is not limited to obvious mistakes. Subtle errors can slip into documents that go through multiple review stages. People tend to trust text that looks professional and reads smoothly. Generative AI produces text that looks professional by design. That smooth surface can mask underlying inaccuracies.

Organizations that adopt these tools without building a verification habit expose themselves to reputational and operational harm. The solution is not to avoid the technology. The solution is to understand its weaknesses and build checks into your workflow.

Practical Techniques for Checking Outputs

You do not need to be a technical expert to verify what an AI generates. A few straightforward habits can catch the majority of errors before they cause trouble.

Verify Specific Claims

Treat every specific fact in the output as a hypothesis until you confirm it. Look up dates, names, numbers, and quotes separately. If the model cites a source, find that source yourself. If the source does not exist, discard that part of the output.

Ask the Model to Explain Its Reasoning

Some models can produce a step-by-step explanation of how they arrived at an answer. This chain of thought often reveals errors. If the model says it used a formula but applies it incorrectly, the reasoning will show the mismatch. You can then correct the output or reject it.

Cross-Check with a Second Query

Ask the same question in a different way. Rephrasing can cause the model to generate a different response. If the two answers contradict each other, you have a clear sign that at least one of them is wrong. Compare the two and look for consistent facts.

Set a Lower Temperature for Factual Tasks

If the tool you use allows parameter control, a lower temperature setting makes the model more deterministic and less creative. This reduces the chance of the model inventing details. Higher temperature settings are useful for brainstorming, but they increase the risk of hallucination.

Use Explicit Instructions in Your Prompt

Tell the model to only answer when it is certain and to say "I do not know" otherwise. Some models respond to this instruction and produce fewer invented facts. You can also ask the model to provide citations from a known source or to limit its answer to information from a specific document.

These techniques are not foolproof. They reduce risk but do not eliminate it. The best practice is to treat every AI output as a draft that needs human review, not as final work.

How Prompt Design Affects Accuracy

The way you phrase your request has a direct impact on the quality of the response. Vague or open-ended prompts give the model more room to invent. Specific prompts with clear constraints produce more reliable results.

For example, instead of asking "Tell me about the history of Singapore," try "List five key events in Singapore's history between 1965 and 1990, with the year of each event." The second prompt limits the scope and forces the model to produce specific, checkable facts. If the model invents an event, you can spot it quickly because you asked for something verifiable.

Including context in your prompt also helps. If you are asking about a specific document or dataset, describe it briefly. The model can use that context to narrow its predictions. Without context, the model draws on its broad training, which increases the chance of mixing up unrelated information.

What This Means for Learning and Development

For teams that create training materials, process documentation, or internal knowledge bases, the risk of spreading a hallucinated fact is especially high. A single error in a training guide can be repeated by dozens of people before anyone notices. The cost of correcting that error later is much higher than catching it early.

One practical approach is to build a verification step into the content creation workflow. After an AI generates a draft, have a human reviewer check every claim against a trusted source. Over time, the team develops a sense of which types of queries produce unreliable output and can adjust their prompts accordingly.

Another useful habit is to keep a log of corrections. When a hallucination is caught and fixed, note what the original prompt was and what the correct answer should be. This log helps the team learn which topics or phrasings are risky and can inform future prompt design.

Wrapping Up: Trust but Verify

Generative AI is a powerful tool for boosting productivity and sparking ideas. But it is not a source of truth. The same mechanism that makes it fluent also makes it prone to creating plausible falsehoods. Understanding this trade-off is the first step to using the technology responsibly.

By treating every AI output as a first draft, verifying specific claims, and designing prompts that reduce ambiguity, you can catch most errors before they reach your audience. The goal is not to eliminate all risk - that is unrealistic - but to manage it well enough that the benefits of the tool outweigh the occasional correction.

United Ceres College, located at 37 Middle Rd, #05-00, Singapore 188946, offers workforce learning programs that help professionals build practical skills for working with generative AI. You can reach the college at +65 8699 3609 for more information.