DIGITAL HUMAN ASSISTANTS

The Reality Gap: How AI Hallucinations and Fabrications Can Impact Your Business

What are Hallucinations: They are when LLMs or MFMs (Multimodal Frontier Models – such as Anthropic's Claude Sonnet 3.5 or OpenAI's o1-preview) provide inaccurate or incomplete information, usually based on existing knowledge but are applied incorrectly.

Hallucinations stem from the fundamental mathematical and logical structure of LLMs. Evidence shows that at every stage of the LLM pipeline - from training data compilation to fact retrieval and intent classification – there is a non-zero probability of hallucination (so they will occur!). It is, therefore, impossible to eliminate them through architectural improvements, dataset enhancements, or fact-checking mechanisms so we need to be vigilant!

Type	Definition	Example
Pedantic Hallucinations	Correct but non-informative content that adds no value.	Repeatedly paraphrasing what was already stated, like defining a tree as "a tall plant with a wooden trunk" when more specific information was needed.
Instruction Inconsistency	Content that violates prompt requirements or goes off-topic.	When asked about climate change in Europe, discussing global weather patterns in general.
Context Inconsistency		
- **Overgeneralization**	Converting individual cases into broader claims.	Taking one person's opinion about a product and presenting it as "customers generally believe...".
- **Oversimplification**	Inappropriately narrowing complex information.	Reducing multiple causes of a financial crisis to just one factor.
Factual Incorrectness	Providing specifically wrong information.	Stating a blood sugar level is 150 mg/dL when it's actually 120 mg/dL.
Misinterpretation Types		
- **Corpus Misinterpretation**	Misunderstanding context within training data.	Confusing historical events from different time periods.
- **Prompt Misinterpretation**	Misunderstanding user intent or question.	Confusing "lead" (chemical element) with "lead" (leadership).
Needle in a Haystack Errors		

Type	Definition	Example
- **Missed Key Data Points**	Omitting crucial information.	Mentioning only one cause of World War I when discussing its origins.
- **Partial Incorrectness**	Mixing correct and incorrect information.	Stating Neil Armstrong walked on the moon in 1959 instead of 1969.

What are Fabrications: They involve creating entirely false information with no basis in the training data.

Type	Definition	Example
Complete Fabrication	Creating information with no basis in the source material.	Generating non-existent scientific studies or research papers.
Subtle Wording Changes	Small alterations that change meaning without obvious contradiction.	Changing "might be effective" to "is effective" in medical claims.
Citation Fabrication	Creating false references or sources.	Generating non-existent legal cases in court filings.
Detail Embellishment	Adding fictional details to factual frameworks.	Creating specific statistics for a true event where no such statistics exist.
Entity Fabrication	Creating non-existent people, organisations, or events.	Generating fake quotes from historical figures.

Key Distinctions

- **Hallucinations** typically involve mishandling existing information.
- **Fabrications** involve creating new, unsupported information.
- Both can appear in subtle forms that are hard to detect so you are your teams need to be aware and vigilant!
- They can result in Misinformation, Misrepresentation, Erosion of Trust, Amplification of Bias and Reputational Harm.
- **Want to learn more about how they can be reduced? Reach out to our team!**

Check out the latest Digital Human Assistants news today!

Remember, you can't outsource your company's risks! This is not about AI malevolence or intentional deception. LLMs do not possess consciousness or intent; they generate content based on patterns in data they were trained on.