Taming AI Hallucinations: Why They Happen, and Whether RAG Can Help

Author: Wenceslao Alfageme
Published: 1st April 2025

Previously, I posed the question: *Does building a second brain really solve our data overload issues, or does it just create new challenges?* , I explored whether building a “second brain” truly helps with information overload, or just adds another layer of complexity.

This time, I want to focus on a key challenge that keeps popping up in my experiments with local LLMs and multi-agent setups: **AI hallucinations**.

These aren’t just amusing peculiarities. They expose a fundamental limitation in how large language models (LLMs) work, and why they can sound so convincing, even when they’re wrong.

So, What Are AI Hallucinations?

To those less familiar with the technical side, “hallucination” might sound like a computational glitch. But as AI researcher Ignacio de Gregorio puts it in "***ChatGPT is Bullshit: The Hallucination Lie***", hallucinations are really a result of how LLMs **generate text**, not from understanding, but from guessing.

These models don’t _know_ things. They rely on probability to predict what word should come next, based on the patterns they’ve seen in huge training datasets. They aren’t checking facts, they’re aiming for fluent, coherent language.

That’s why LLMs can produce statements that _sound_ accurate, but are completely detached from reality. And if we’re not verifying their outputs, that’s where real risks begin.

Why Do LLMs Hallucinate?

There are a few key reasons:

- **Probabilistic Generation** – LLMs choose likely-sounding words, not necessarily correct ones.

- **No Perception of Reality** – They don’t "know" facts or have access to real-time context, they’re just processing text.

- **No Built-In Fact-Checking** – Their goal is linguistic coherence, not truth.

This becomes especially problematic when we're building workflows or systems that rely on AI to summarise research, generate business insights, or assist in decision-making.

Note: To understand how LLMs work, this is one of the best explanations I have found by AI researcher Ignacio de Gregorio here.

Can Retrieval-Augmented Generation (RAG) Help?

A promising workaround is **Retrieval Augmented Generation**, or RAG. It's designed to ground AI responses in actual, external data:

1. The system retrieves information from a trusted knowledge base

2. The LLM uses that context to generate a more accurate response

In theory, this should limit hallucinations, if the model can only pull from vetted sources, it's less likely to fabricate.

But RAG isn't a silver bullet. It still depends on:

- The **quality and freshness** of the source material

- The LLM actually using the retrieved content properly

- Careful system design to avoid performance trade-offs or ignored data

Where Does That Leave Us?

LLMs can be incredibly powerful, but their confidence can be misleading. And when we're building tools like an "AI second brain," those small inaccuracies can ripple into bigger problems.

So the real question is: **How do we design AI systems that account for these flaws while still delivering value?**

Coming Up Next:

In the next article, I'll move from _why_ hallucinations happen to _how_ we can mitigate them, especially through multi-agent workflows and better orchestration models. Because no matter how clever an AI is, it's only as reliable as the process and data behind it.

Would love to hear how others are approaching this. Have you experienced hallucinations in your AI workflows? Have you tried using RAG or multi-agent setups to reduce them? Drop your thoughts in the comments, always keen to learn from others facing similar issues or concerns.

References:

- Previous Article: *In the Age of Information Gluttony*
- Previous Article: *Is an AI Based Second Brain the Best Path? Rethinking AI Solutions with Agile Thinking*
- *Ignacio de Gregorio* Medium Article: *ChatGPT is Bullshit: The Hallucination Lie*

Hashtags: #AI #CrewAI #AgileMethodology #DigitalTransformation #RAG #SecondBrain #LLMs

Definitions

Hallucinations
In the context of AI, “hallucinations” occur when a model (especially a Large Language Model) generates outputs that sound plausible but are factually incorrect. This happens because the AI is predicting words based on patterns in training data, rather than verifying truth.

LLM (Large Language Model)
A type of artificial intelligence trained on vast text datasets to predict the next word in a sequence. Examples include GPT-style models. While powerful, LLMs can produce confident responses that may not always align with real-world facts.

Probability Distribution
A mathematical way of describing how likely each possible outcome (or next word) is. In an LLM, the model calculates probabilities for many potential word choices—some correct, others incorrect.

RAG (Retrieval Augmented Generation)
A technique where an AI model retrieves specific, curated information from an external source before generating a final response. This grounding can help reduce hallucinations, though it requires careful maintenance of the underlying data and systems.

Training Data
The text or information used to “teach” an AI model. If the training data is incomplete or flawed, the AI’s outputs will reflect those gaps or biases, leading to hallucinations or inaccuracies.

Knowledge Base
A collection of verified facts, documents, or other data sources that an AI can reference. When used with RAG, it aims to provide more accurate, evidence-based responses.

Disclaimer and Disclosure

Third-party Content and AI Assistance: This article references tools and software that are publicly available and proprietary to their respective creators. The author does not claim ownership or affiliation with these third-party products. This article was written by the author with assistance from Generative AI Language Models.
Transparency Notice: While every effort has been made to ensure accuracy, readers should verify information independently and consult official sources or documentation for the mentioned tools and software. The use of AI in the writing process is disclosed in the interest of transparency, but all opinions and analyses are the author’s own unless otherwise stated.