SHIFT

SHIFT TECHNOLOGY **INSURANCE PERSPECTIVES**

THE AI-HALLUCINATIONS EDITION

www.shift-technology.com

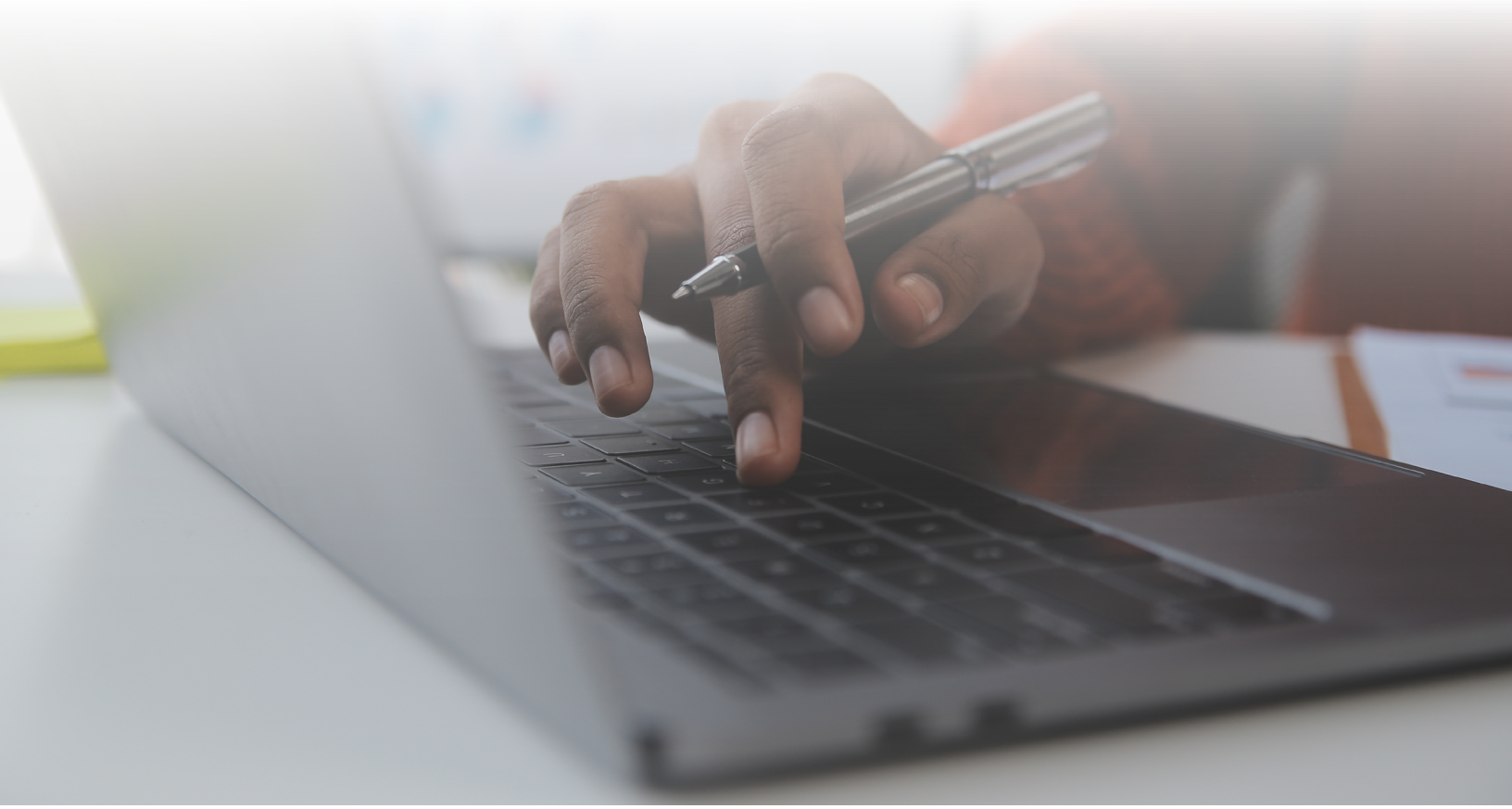
From the editor

The increasing use of generative AI (Gen AI) "out in the wild" is driving mainstream awareness of the problem of hallucinations. This describes a situation where a language model answers a prompt with full confidence, only for the user to find that under further scrutiny, significant portions of the response are partially or fully false.

When asking for restaurant recommendations, looking to settle a bar bet, or accessing some other innocuous bit of information these "AI hallucinations" might be fascinating, funny, or frustrating (or all three...). In some cases, for example trying to identify if a specific type of wild mushroom is poisonous, hallucinations could even be dangerous. In the realm of insurance AI, the ability to mitigate the risk of hallucinations could mean the difference between the success or failure of AI-based initiatives such as claims automation or fraud detection.

But to understand how to stop AI hallucinations we need to know exactly what they are; the mechanisms behind their emergence; and why models seem unable to say "I don't know" instead of simply making up an answer.

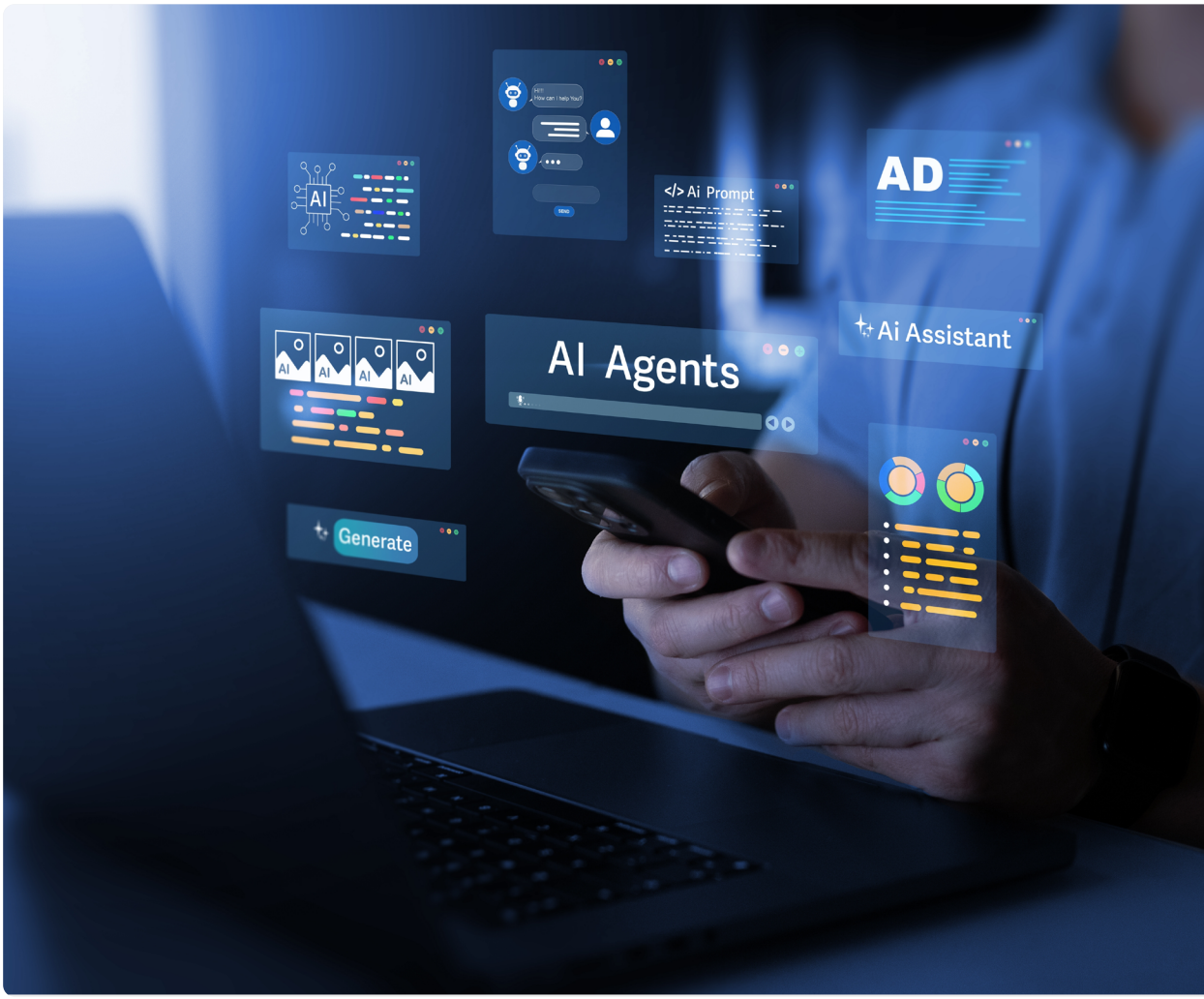
In this edition of "Shift Insurance Perspectives" we will explore why making models more humble is just as important as making them smarter and what that means for the future of insurance AI.



Defining AI hallucination

When we say that a Large Language Model (LLM) "hallucinates" we are not attributing a human condition to the AI or implying that the AI is acting under the influence of an outside force such as malware. By "hallucinates" we mean that based on the prompt provided the model produces confident, fluent answers that are factually wrong. In certain cases, the LLM may recognize similarities between an input and something it experienced several times and embrace that pattern, totally deviating from the case initially proposed.

For insurers adopting AI to support their processes, these errors can range from relatively harmless (an out-of-context citation on a greeting message) to critical (a wrong policy clause or a made-up field in a claim's document, among others). What is important to remember however is that the problem is not that models make mistakes, it is how they make them: confidently, coherently, and without signaling uncertainty.



Where do hallucinations come from

To understand where hallucinations come from we must first understand how language models work. Whether large or small, language models are fundamentally probabilistic next-word predictors. They are not truth engines. They learn to produce the most likely continuation of a text given a specific context. Models do not, by their design, check that the continuation delivered is true.

They learn to produce the most likely continuation of a text given a specific context. Models do not, by their design, check that the continuation delivered is true.

What this means is that an LLM generates text in a way that's surprisingly similar to how a phone's predictive keyboard works: it looks at the previous words and guesses what should come next. We can think about the process this way: a user types "see you" on their smartphone and simply accepts the first suggestion that pops up. The intended message may have been "see you later, pal," but since you most recently texted your significant other as opposed to your poker buddy, you send, "see you later, my love" instead.

And while this example may seem a little simplistic, it is generally how a language model works. LLMs just happen to operate on a massive scale and are trained on billions of sources instead of an individual's recent texts.

With this comparison in mind, it can be easier to understand why hallucinations emerge:

- **Training focus:** The model is rewarded for sounding plausible, not for being accurate. This focus minimizes the difference between the model's predictions and training examples, but those examples don't include "I don't know" cases.
- **Incomplete or ambiguous data:** If the training set lacks evidence for a fact or the information given in the prompt is ambiguous, the model fills in the blanks. We can equate this to a clever student guessing an answer during an exam.
- **Prompt pressure:** Users often willingly or unwillingly force a response, especially in a chatbot configuration. Without a built-in mechanism to abstain, the model must respond somehow, even when unsure. In the below example there is no value for C in the file and the value Z never appears.
    - user: "Extract the values A, B, and C from the attached file"
    - LLM: "A=X, B=Y"
    - user: "What about C?!!?"
    - LLM: "Oooh sorry, C=Z"

Understanding the basic mechanism behind how language models work is critical to understanding how to mitigate the risk of hallucinations.

Allowing LLMs to say, "I don't know"

Teaching models to abstain when they are uncertain is a critical aspect in avoiding hallucinations. LLMs must be able to respond with something along the lines of "I don't have enough information to answer that," instead of defaulting to a fabricated answer.

This concept relies on the introduction of an abstention option along with the correct and all of the incorrect answers. Technically, this requires models (and their evaluation benchmarks) to support uncertainty estimation, calibrated confidence scores, or abstention thresholds.

What this approach delivers are models that could choose to say nothing unless its internal confidence exceeds a certain level.

Interestingly, the idea of abstention in insurance AI is not entirely new. Shift has long used abstention to control the behavior of our machine learning models, for example document classification algorithms. Instead of always choosing the top class (say, "invoice," "ticket," or "ID card"), models were allowed to abstain when a confidence score was below an established threshold. The company has continued making abstention a critical aspect of its work in Gen AI and agentic AI.

Conclusion

AI is bringing tremendous value to the insurance industry by supporting critical initiatives including claims automation, fraud, subrogation, and underwriting risk detection, and many other processes. However, it must be clearly understood that AI hallucinations pose a threat to ongoing success. Insurance AI must be given the ability to say, "I don't know." It must be given the option to default to the human in the loop. It must not be forced to give an answer in which it is not certain. LLM abstention is a crucial aspect of insurance AI, and as important, a critical element to more responsible AI.

SHIFT