RT ProExec

Why AI Hallucinates, and Why That's Now an Insurable Event

CONTACT
RT ProExec
rtproexecinfo@rtspecialty.com

Or contact your local RT ProExec broker at rtspecialty.com

THE MAIN POINT

AI hallucinations are persistent and non-negligible. Researchers at OpenAI recently published a paper explaining that hallucinations arise from how these systems are trained and evaluated: models are rewarded for producing answers, even when they are uncertain, which incentivizes guessing. That means hallucinations are not something that can be fully engineered away in the near term.

Your clients are using these tools to write contracts, screen resumes, support clinical decisions, answer customers, and generate code. When the AI confidently delivers an incorrect or incomplete answer and that information is relied upon, it can create legal and financial exposure.

That raises a fundamental insurance question: how should these risks be evaluated and allocated? The market is beginning to develop clearer answers.

Why AI Gets It Wrong (in Plain English)

A September 2025 paper from OpenAI and Georgia Tech titled "Why Language Models Hallucinate" breaks it down[1]. You do not need a PhD to follow it. Here is the gist.

Reason #1: AI is trained and evaluated like a student cramming for a multiple-choice test.
Think back to the SAT. If you did not know an answer, what did you do? You guessed. A blank got you zero, and a guess at least had a shot. That is essentially how AI models are graded during training. A confident guess has a chance at earning a point. "I don't know" earns zero. So, the model learns to do exactly what students do: guess. We call those guesses hallucinations.

The researchers put it bluntly: these models are effectively always in test-taking mode, because the way we score them rewards guessing over acknowledging a lack of knowledge.

Reason #2: The AI does not actually "know" things in the way users expect.
When you ask the AI a question, it is not looking up an answer. It is predicting which words should come next based on patterns. If your client asks for "the case that established X legal principle in Florida," the model may produce something that looks like a citation, even if no such case exists. It is not intentionally misleading, but when information is rare, highly specific, or weakly represented in training data, the likelihood of error increases. The more specialized the question, the higher the risk of hallucination.

Reason #3: Adding search and "reasoning" does not fix it.
Vendors love to say "we use retrieval" or "our model reasons" as if that solves hallucination. These approaches can improve accuracy, but they do not fully resolve hallucinations. The paper addresses this directly: when search fails to find the right answer, the model still guesses, because guessing is still what it was trained to do. Newer reasoning models are better at some tasks, but they can still fabricate citations, case law, and clinical guidance.

The bottom line for anyone deploying AI: it will be wrong, confidently, some percentage of the time, and that risk cannot, at least today, be fully engineered away. The practical question is how that risk is managed.

So, What Does This Have to Do With Insurance?

Everything.
For the last few years, AI claims have been fought out under General Liability, Technology E&O, Cyber, and Media policies. Carriers paid some, denied others, and many are now drawing clear lines. At renewal, a growing number of traditional policies are adding express AI exclusions, some of them broad, some described by their carriers as absolute.

The risk has not disappeared. It is increasingly being carved out of certain policies and shifting toward a developing specialty market. That market is still developing, but it is real, and it is where many of us as wholesale specialists are now spending a meaningful part of the day.

Rather than focus on a single carrier, this piece walks through how we are typically seeing standalone AI liability coverage structured across the specialty market today, including common. It covers the kinds of coverage parts, the typical conditions, and certain exclusions that can ultimately determine whether a claim is covered..

What Standalone AI Liability Coverage Typically Includes

The standalone AI liability forms emerging in the specialty markets tend to be built around a set of distinct coverage parts. Each one targets a different way an AI system can cause harm to a third party. Exact wording varies by carrier, but the categories appear to be converging on the following:

1. AI Error. Responds when the AI produces a wrong answer that causes a third-party financial loss. The classic example: an AI tool generates flawed financial analysis, legal research, or code, and the party who relied on it suffers a loss. This is the AI version of a professional-services error.

2. AI Intellectual Property Infringement or Defamation. Responds when an AI output defames a person or business, or infringes a copyright or trademark, for example AI-generated marketing copy that disparages a competitor or reproduces protected material. One thing to keep in mind however: across most forms in this market, patents and trade secrets are treated separately and are commonly excluded from this part.

3. AI Unauthorized Data Disclosure (AI Data Leakage). Responds when an AI system unintentionally exposes someone's protected information in its output, for example a chatbot that is manipulated into revealing personal information contained in its training or context data.

4. AI Bodily Injury. Responds when a person is physically harmed because an AI system delivered the wrong information or instruction.

5. AI Property Damage. Responds when an AI system causes tangible damage to physical property.

Common Coverage Conditions

Standalone AI liability commonly carries a set of conditions the insured must satisfy before coverage will respond. Miss one, and the claim may not be paid. The conditions you will most often see across the market include:

- **A written AI usage policy.** The insured maintains an internal policy governing what employees can and cannot do with AI tools.
- **Alignment to a recognized AI governance framework.** The insured deploys AI in line with a nationally or internationally recognized risk management framework, most commonly the NIST AI Risk Management Framework[2] (including its Generative AI Profile) and/or ISO / IEC 42001[3]. If a client has never heard of NIST AI RMF, that is a gap to close before submission. (We can help with that of course.)
- **No tampering with vendor guardrails.** The insured does not strip out or disable the AI vendor's built-in safety features. Jailbreaking your own system can avoid coverage.
- **No unapproved retraining or fine-tuning.** This would be considered, and what is repeatedly said in many AI policy forums, as the scheduled use case for the AI solution itself. Coverage generally applies to the AI as deployed, not to a system the insured has materially modified through retraining or fine-tuning that changes its functionality or performance.
- **Disclosure to end users.** The insured clearly discloses to third parties, at or before the point of interaction, that they are engaging with or relying on AI-generated output. Using AI on customers without telling them can become a coverage problem.
- **Reviewed privacy notices.** Where the AI touches protected information, the privacy notices have been reviewed by a qualified legal professional.

These are important points to review against the specific form at issue and to discuss with the client prior to binding, as they may impact how coverage responds.

Approaching AI Risk in Practice

Your clients are likely using AI, and it is possible that an exclusion relating to AI may be added at renewal. A functioning specialty market is already quoting real standalone AI liability coverage, including capacity from large carriers and major reinsurer paper. The submissions are technical, and the underwriting conditions are real. Consider talking with your clients deploying AI about their governance and controls. Insurers often look at factors such as formal usage policies, alignment with recognized frameworks, and human oversight when evaluating coverage. For clients that develop or sell AI-enabled products, it can also be helpful to understand how those offerings are viewed from an insurance standpoint, as procurement teams are increasingly asking whether solutions are insurable and whether vendors carry coverage to support their indemnity obligations. If helpful, consider working with a wholesale broker familiar with this market.

The Bottom Line

The researchers' conclusion is that sobering hallucinations are not going to be engineered away using current methodology, because the incentives that produce them are baked into how AI is built. The field may eventually change how it grades models. In the meantime, math says these systems will keep being confidently wrong at least part of the time.

That reality has clear insurance implications. As organizations deploy AI more broadly, questions around whether existing policies respond, where exclusions apply, and when specialized coverage may be appropriate are becoming harder to ignore.

Structuring coverage for AI exposures can be nuanced. Our team at RT Specialty works with retail brokers to support the submission process and facilitate placement in the market. If you have a client deploying AI and are looking for support in navigating the submission, we would be happy to connect.

SOURCES AND CITATIONS

1. Kalai, A.T., Nachum, O., Vempala, S.S., and Zhang, E. (2025). "Why Language Models Hallucinate." arXiv preprint arXiv:2509.04664, September 4, 2025. OpenAI and Georgia Institute of Technology. https://arxiv.org/abs/2509.04664

2. National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework: Generative AI Profile (NIST AI 600-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

3. ISO / IEC 42001:2023. Information technology, Artificial intelligence, Management system. https://www.onetrust.com/resources/navigating-the-iso-42001-framework-ebook/?ef_id=Cj0KCQjw_vnQBhCxARIsADcZyxIgqgTK9Dxat3KdcRD-sqxgEZ00bjvDJmyPuBrzRuBkF-T0jNZMwWIaAv0bEALw_wcB:G:s&s_kwcid=AL!17820!3!764120771997!e!!g!!iso%2042001!21957929477!171259788183&utm_source=google&utm_medium=cpc&utm_campaign=G|NA|Search|Non-Brand|Data_AIGovernance|US&utm_content=Resources_ISO_42001&utm_term=iso%2042001&gad_source=1&gad_campaignid=21957929477&gbraid=0AAAADCQCNdtCQggjTrFTYjtvCXKD6O62&gclid=Cj0KCQjw_vnQBhCxARIsADcZyxIgqgTK9Dxat3KdcRD-sqxgEZ00bjvDJmyPuBrzRuBkF-T0jNZMwWIaAv0bEALw_wcB

All sources were accessed between March 11 and April 2, 2026. This article does not reproduce copyrighted material; all factual claims are paraphrased from the cited sources. Specific companies referenced in the underlying source material have been anonymized where appropriate.