AI Hallucinations in Drug Discovery: Examples & Detection

By Adrien Laurent, CEO at IntuitionLabs • 3/12/2026 • 30 min read

ai hallucinations | drug discovery | large language models | pharma research | machine learning | ai detection

AI Hallucinations in Drug Discovery: Real Examples and How to Catch Them

Executive Summary: Artificial intelligence (AI) and large language models (LLMs) promise to accelerate and transform drug discovery, but a critical vulnerability has emerged: *hallucinations*. These are cases where AI systems generate plausible but incorrect or non-existent information. In drug discovery, hallucinations have led to fabricated scientific citations, invented disease mechanisms, and false compound proposals. For example, ChatGPT produced completely fake PubMed references with mismatched PMIDs ([1] pmc.ncbi.nlm.nih.gov), and an AI model listed nonexistent medications as interacting with an herb (www.amazon.science). Such errors can mislead research and endanger patient safety in high-stakes pharmaceutical contexts ([2] www.sinequa.com) ([3] blog.biostrand.ai). To date, addressing hallucinations remains difficult. No foolproof, automated detector exists ([4] www.drugdiscoveryonline.com). However, promising mitigation strategies are emerging. These include rigorous post-generation verification (comparing AI outputs against validated databases or experimental data ([5] blog.biostrand.ai) ([6] pmc.ncbi.nlm.nih.gov)), retrieval-augmented generation (grounding AI responses in real documents ([7] pmc.ncbi.nlm.nih.gov) ([8] www.sinequa.com)), prompt engineering and chain-of-thought prompting ([9] blog.biostrand.ai) ([10] www.drugdiscoveryonline.com), fine-tuning on curated biomedical data ([11] blog.biostrand.ai), and human-in-the-loop review ([12] www.drugdiscoveryonline.com) ([6] pmc.ncbi.nlm.nih.gov). Furthermore, new tools like “HalluMeasure” decompose AI-generated text into individual claims to check factual validity (www.amazon.science). Regulatory bodies are also taking note: the FDA has proposed guidance for verifying the **credibility** of AI models used in drug development ([13] www.fda.gov). This report systematically examines the origins, examples, and impacts of AI hallucinations in pharmaceutical R&D, reviews detection and mitigation techniques with evidence, and discusses future directions. Our analysis underscores that while AI hallucinations have catalyzed innovative ideas in some views ([14] kingy.ai), in drug discovery their risks currently demand vigilant human oversight and robust validation frameworks.

Introduction and Background

AI has become deeply integrated into pharmaceutical research and development. In silico methods, from quantitative structure–activity relationship (QSAR) models to modern deep learning, now support target identification, lead optimization, and clinical trial design. Advances in neural networks and LLMs have brought unprecedented capabilities: language models can summarize millions of biomedical papers, propose novel molecular structures, and even suggest clinical trial protocols. As one review notes, AI “holds significant potential” to transform drug discovery at every stage—uncovering disease mechanisms, generating novel drug candidates, and optimizing trials ([15] pmc.ncbi.nlm.nih.gov) ([16] pmc.ncbi.nlm.nih.gov). Even modest improvements matter: Morgan Stanley projected that a slight AI-enabled boost in early-stage success rates could yield an additional 50 novel therapies over 10 years (≈$50 billion in value) ([17] www.drugdiscoveryonline.com).

However, this optimism comes with caveats. Early discussions cautioned that biological and chemical data differ fundamentally from image or language data, meaning AI successes in vision do not automatically scale to drug discovery ([18] pmc.ncbi.nlm.nih.gov) ([19] pmc.ncbi.nlm.nih.gov). As AI models tackle increasingly complex tasks, the risk of **hallucination**—confidently generating false information—has become a major concern. Hallucinations in AI are analogous to a clinician confidently applying an incorrect theory: they can lead AI to fabricate plausible-sounding but incorrect facts, such as nonexistent molecular scaffolds, spurious clinical trial suggestions, or invented literature citations. In a pharmaceutical context—where errors can misallocate millions of R&D dollars or compromise patient safety ([2] www.sinequa.com) ([3] blog.biostrand.ai)—the stakes are particularly high. The term “hallucination” has thus become widespread among AI practitioners to describe this phenomenon of AI “making things up” ([20] www.drugdiscoveryonline.com) ([21] www.pharmavoice.com).

Hallucinations arise because generative AI models learn statistical patterns from data rather than true underlying principles. State-of-the-art LLMs predict the next token or phrase given text prompts, leveraging massive pretraining. They have no built-in factual database or reasoning engine, so when prompted beyond their knowledge, they interpolate based on spurious patterns. Even minor "creative" errors can be dangerous. As Di Gioia and Foppen explain, LLMs can generate outputs that "are incorrect, misleading, or nonsensical" with "high degree of confidence" ([20] www.drugdiscoveryonline.com). An AI answer that "sounds very confident" may not be true—a pitfall poignantly noted: "What these LLMs are designed to do is produce really good English and convince you that's the truth" ([22] www.pharmavoice.com). This report analyzes how such hallucinations manifest in drug discovery, surveys real examples from the literature and industry, and evaluates strategies to detect and mitigate these errors.We draw on academic studies, industry white papers, and expert commentary to provide an in-depth, evidence-backed picture of the problem and its solutions.

The Nature of AI Hallucinations

Definition: AI hallucination typically refers to a model output that is linguistically coherent but factually incorrect or unsupported by data. It is not merely a trivial error but an error delivered as plausible truth. Common definitions include: *"the generation of content that is not based on real or existing data but produced by a model's extrapolation or creative interpretation"* ([21] www.pharmavoice.com), and outputs that "do not correspond to the reality of the input data" ([23] www.drugdiscoveryonline.com). In other words, hallucinations are model outputs that "seem plausible but are verifiably incorrect" (www.amazon.science). They effectively fabricate information – invented references, mechanisms, or entities – rather than retrieve or compute known answers.

Why they occur: Hallucinations stem from fundamental properties of generative models. Large language models (LLMs) are trained on broad internet text using a probability-based next-word objective ([24] blog.biostrand.ai). They excel at learning correlations in language but lack genuine understanding. When tackling a question, an LLM does not "lookup" facts; instead it **predicts** text that statistically fits the prompt. If the training data lacks a precise answer or contains noisy information, the model still outputs something plausible-sounding. As Salinas et al. (Amazon) note, LLMs will predict a mix of real and fictional medications when asked about drug interactions, because they "don't search a medically validated list… [but] generate a list based on the distribution of words associated with St. John's wort" (www.amazon.science). In short, hallucinations arise because: (a) the model's training data is incomplete or contains errors, and (b) the model has no built-in mechanism to verify facts.

Technical factors exacerbate this. Hallucinations are more common in out-of-distribution queries or rare facts. One study observes: *"LLMs will hallucinate at least the same percentage of rare facts as appear only once in their training."* For example, if 15% of chemical property mentions were unique, expect ~15% hallucination on queries about them ([25] www.linkedin.com). Moreover, the search space for drug-like molecules or biomedical knowledge is enormous; limited training coverage means models often guess. Prompt techniques also matter: poorly constrained or overly open prompts give the model "invitation" to guess beyond its knowledge.

Frequency and Risk: Hallucinations are not a fringe issue. A preprint from Biostrand (2024) cites a public "hallucination leaderboard" showing top LLMs with hallucination rates ~3% and lower-ranked models up to 27% ([26] blog.biostrand.ai). In one experiment, GPT variants produced fake reference citations 29–33% of the time ([26] blog.biostrand.ai). Even at the low end (~3%), in drug discovery that is unacceptable: a 3% rate could compromise critical decisions. As one analyst notes, AI systems with as low as 5% hallucination become "completely unusable for clinical decision support" ([27] www.linkedin.com). In fact, the Biostrand authors state that even a 3% hallucination rate "can have serious consequences in critical drug discovery applications" ([28] blog.biostrand.ai). Thus, hallucination frequency, even if seemingly small, translates to significant risk in the pharma domain.

Hallucinations come in several **flavors**. One can distinguish (though terminology varies):

- *Factual Hallucinations:* Inventing false facts, such as a nonexistent research paper or compound.

- *Logical/Contextual Hallucinations:* Generating outputs that contradict known context or self-inconsistencies.
- *Structural Hallucinations:* For generative models that propose chemical structures, outputting an invalid or impossible molecule.

In drug discovery, examples range from the first (false references) to the last (invalid molecule proposals). Crucially, hallucinations are not just trivia mistakes: they can misset entire research directions. For instance, an LLM might confidently propose an invalid medicinal chemistry pathway that leads researchers astray, or a predictive model might forecast an incorrect toxicity profile due to spurious correlations in its training data. These issues highlight the need for domain-specific vigilance: unlike a chatbot in harmless banter, an AI in pharma must be anchored very closely to factual reality.

Table 1 below summarizes notable examples of AI hallucinations affecting biomedical and pharmaceutical contexts, illustrating the diverse manifestations of the problem. Each entry reflects a documented case or study, with the potential impact on drug discovery or medical decision-making.

Example	Context / Model	Hallucination	Impact	Source
Fabricated references in scientific writing	ChatGPT (GPT-3.5) answering research inquiries	Invented journal articles and PMIDs unrelated to query	Misleading literature reviews, wasted verification effort	[23†L121-L129] [20†L105-L113]
Nonexistent disease association	ChatGPT (GPT-4) on metabolic disorder question	Described liver involvement in LOPD (not supported by any data)	Could spur false research leads	[23†L145-L153]
Fictional drug interactions list	ChatGPT answering "interactions with St. John's wort"	Mixed real meds with completely fictitious drug names	Danger of advising erroneous treatments	[35†L39-L44]
Incorrect chemical mechanisms	ChatGPT in organic reaction explanations (education setting)	Mechanisms with one or more incorrect steps (while sounding fluent)	Miseducation or misguidance in chemistry reasoning	[44†L13-L16]
Overconfident medical summary	Early chat model summarizing patient records (PharmaVoice account)	Accurate fluency but occasional factual mistakes	Potential misdirected clinical decisions	[30†L38-L42]
RAG-pipelined vulnerability (Anticocaine study)	GPT-4 assisting in drug design	General suggestions, but authors note GPT-4 still "liable to false narratives"	Emphasizes need for expert validation	[41†L271-L279]

Table 1: *Reported examples of AI hallucinations in biomedical/pharmaceutical settings. Each illustrates plausible-sounding but false outputs, with consequences for research or patient care.* In some cases (left two rows), hallucinations occur during literature search or reasoning, while others involve education or drug interaction tasks. Across all, the solution involves cross-checking: experts manually verified GPT-4 outputs in the anticocaine study ([6] pmc.ncbi.nlm.nih.gov) and compared ChatGPT's citations to PubMed, finding 15% were fake ([29] pmc.ncbi.nlm.nih.gov). We now examine these and related examples in detail.

Hallucinations in Drug Discovery: Causes and Consequences

Drug discovery poses unique challenges that exacerbate AI hallucinations. First, the underlying data is complex and often scarce. As Bender and colleagues note, even large chemical datasets are tiny compared to image data; inherently novel compounds and targets may lack precedent in the training corpus ([30] pmc.ncbi.nlm.nih.gov) ([31] pmc.ncbi.nlm.nih.gov). Second, decisions in pharma are high-stakes: a synthetic chemist expending resources on an AI-suggested molecule, or a clinician considering an AI-suggested therapy, cannot afford unchecked errors. Hallucinated data in drug discovery can thus lead not just to wasted effort but to direct patient harm.

For example, an AI hallucination in lead generation could propose a "novel" molecule that violates known chemical rules (e.g. impossible valences, unstable motifs). Using such output as a real candidate could scramble lab resources. Likewise, an AI summary that misstates a drug's side effects could mislead clinical trial design. The sine qua non is that in this domain, "outputs not grounded in the organization's actual evidence base…is not just an efficiency problem — it is a compliance risk that can derail an approval" ([2] www.sinequa.com). In short, whereas a hallucination from a social chatbot is inconvenient, a hallucination from an AI in drug discovery can have far-reaching adverse impacts.

Regulatory and business implications: The industry is grappling with this. Pharma companies and regulators are now explicitly addressing AI credibility. The FDA's 2025 draft guidance emphasizes that for any AI used in drug approvals, "model credibility" must be **demonstrated** through a risk-based framework ([13] www.fda.gov). In parallel, experts warn that overstated trust could "cause damage to whole industry" by eroding confidence ([32] www.pharmavoice.com). On the business side, hallucinations threaten the return on investment in AI. A hallucinated data entry in a regulatory document could delay a new drug by months. Efforts to catch such errors add cost. By contrast, preventing hallucinations (for example via rigorous knowledge grounding) can maintain trust and accelerate development. The economic opportunity is large — any improvement is valuable — but only if reliability is assured.

Real-World Examples of Hallucinations

We now detail concrete instances where AI hallucinations have emerged in scientific or pharmaceutical contexts. These cases illustrate the different forms hallucinations can take.

ChatGPT Fabricated References in Scientific Essays

A striking example is from multiple recent studies where ChatGPT was asked to write scholarly text with citations. In stem cell research and other biomedical domains, analysts found that a significant fraction of ChatGPT's references were entirely **fabricated**. In one empirical study of 86 GPT-3.5-generated references, about **15.1%** were entirely fictitious and another 9.3% were erroneous ([29] pmc.ncbi.nlm.nih.gov) ([33] pmc.ncbi.nlm.nih.gov). For instance, ChatGPT fabricated a reference "Kallajoki M, Homocysteine and bone metabolism. *Osteoporos Int.* 2002;13(10):822–7. PMID: 12352394." Searching the given PMID led to an unrelated article on urological surgery ([1] pmc.ncbi.nlm.nih.gov). Upon re-prompting for newer references, the model merely changed years and recirculated the same fakes. A human reviewer lamented that the "titles provided did not exist" and that PMIDs corresponded to unrelated studies ([1] pmc.ncbi.nlm.nih.gov). These hallucinated citations undermine scientific integrity; an unwary reader might cite them in error, polluting the literature.

The mechanism is clear: ChatGPT attempts to sound scholarly by mimicking citation formats but lacks database lookup, so it guesses authors, titles, and PMIDs. The resulting citations are superficially plausible ("Osteoporos Int.", etc.) but completely invented or mismatched. Detecting such hallucinations requires examining each citation: indeed, the cited study manually checked each GPT reference against PubMed, revealing the tallies above ([29] pmc.ncbi.nlm.nih.gov). This has prompted calls for caution. Even ChatGPT itself advises that "caution should be exercised when relying solely on [its] output for factual or authoritative information" ([34] pmc.ncbi.nlm.nih.gov). In a drug discovery context, relying on such text summaries without verification would be dangerous.

Invented Disease-Drug Associations

The hallucination problem extends beyond text to scientific content. For example, in a test of medical writing, ChatGPT was asked about liver involvement in late-onset Pompe disease (LOPD). Pompe disease rarely affects the liver in its adult form, yet **ChatGPT confidently wrote a plausible-sounding essay linking LOPD to liver dysfunction** ([35] pmc.ncbi.nlm.nih.gov). It fabricated a narrative of "liver involvement" that has not been reported in the literature (at least in English). This again was a "verifiably incorrect" output; the authors remarked that in reality, such reports do not exist. An AI generating false medical mechanisms could mislead researchers exploring new therapeutic angles or biomarkers. The only safeguard is subject-matter expertise: the authors emphasize researchers must "verify the accuracy and reliability of responses from ChatGPT using their expertise" ([6] pmc.ncbi.nlm.nih.gov). They treated ChatGPT as a brainstorming assistant rather than an authority, accepting only suggestions that matched vetted knowledge ([6] pmc.ncbi.nlm.nih.gov).

Fabricated Chemical Knowledge in Education

Even in educational settings, LLMs have exhibited hallucination patterns. In one study evaluating ChatGPT's explanations of organic reaction mechanisms, only **28%** of the model's generated mechanisms were fully correct ([36] pubs.acs.org). The rest contained mistakes (incorrect steps or missing reagents) though written with high fluency. For example, ChatGPT might describe a plausible arrow-pushing sequence that thus appears legitimate, yet an organic chemist can easily spot the flaw. The risk here is pedagogical: students relying on AI explanations without checking could internalize misconceptions. In drug discovery, analogous errors might misguide chemists designing synthetic routes or interpreting assay mechanisms. The key lesson is that fluent language does not guarantee factual accuracy; nearly three-quarters of the explanations in the study had at least some error ([36] pubs.acs.org).

Hallucinations in Drug Interaction Queries

AI hallucinations can directly endanger patient care. In a demonstration by Amazon scientists, asking an LLM about drug interactions with St. John's wort produced a list mixing real medications with **fictional ones** (www.amazon.science). Here the model "hallucinates" nonexistent drugs due to statistical word association. If such outputs were taken as genuine, a clinician might avoid or prescribe an irrelevant drug based on a non-existent interaction. Even if some hallucinations seem harmless (e.g. a fictional drug with no real effect), others could cause confusion in pharmacovigilance or clinical decision support. This example underscores that in sensitive domains, even a few percent of hallucinated entries can be dangerous. For the hallucinated items in the LLM output, only careful retrieval of a knowledge base or expert review would catch the fakes. Amazon's blog concludes that identifying and measuring hallucinations is "key to the safe use of generative AI" in medicine (www.amazon.science).

AI as a Creative Catalyst (Controversial View)

Not all commentary treats hallucinations solely as liabilities. Some researchers argue that these "confident wrong answers" might paradoxically enable novel discovery. For example, Pagayon suggests that AI's "creative errors" could lead scientists to entirely new chemical space. The idea is that a model might propose an off-the-wall molecule or target that no human would initially consider; even if improbable, it could spark an unexpected line of inquiry ([14] kingy.ai). In this view, hallucinations are like serendipity—chance novel ideas arising from a vast information space. The Kingy article observes, "recent studies suggest these hallucinations—once considered a liability—could unlock novel solutions in drug discovery" ([14] kingy.ai). Indeed, generative AI has already generated many new compounds; some incidental false positives might still inspire real therapeutic candidates. This optimistic angle argues that some controlled risk-taking ("hallucination as acceptable business cost" ([37] www.drugdiscoveryonline.com)) can be tolerated in early-stage R&D if it yields breakthroughs.

However, even proponents note that any hallucinated idea must be rigorously validated in reality. There are trade-offs: investigating every implausible AI idea is infeasible. As one expert said, real constraints (like medical definitions) still need to be embedded ("world model") in the AI to check outputs ([38] www.pharmavoice.com). The consensus remains that if hallucinations are treated as **ideas to test** rather than facts, they may accelerate innovation. But until AI can flag its own uncertainty, drug developers must assume hallucinations can mislead and build verification into their workflows.

Table 1 above catalogs some of these examples, indicating both "bad" outcomes (wasted effort, risk) and the (contentious) "potentially good" angle of novelty generation. The weight of evidence suggests that, in practice, hallucinations have caused significant concern in pharma circles and are being met with a variety of countermeasures, which we discuss next.

Detecting and Mitigating Hallucinations

Given the risks, deploying AI in drug discovery demands robust methods to **catch** hallucinations before they cause harm. Here we survey technical and process-oriented strategies, organized into (1) model-and-algorithmic approaches, (2)

post-generation verification, and (3) human-in-the-loop safeguards. Table 2 summarizes key strategies alongside advantages and references.

Strategy	Description	Example/Tool	Benefits/Limitations	Reference
Retrieval-Augmented Generation (RAG)	Integrate external knowledge (e.g. literature, databases) during generation so outputs cite real sources.	E.g. specialized drug LLMs with literature retrieval.	Greatly improves factual grounding by tying answers to actual documents ([7] pmc.ncbi.nlm.nih.gov) ([8] www.sinequa.com). Requires curated knowledge base and good retrieval; may still omit unseen info.	[28†L1779-L1787] [49†L38-L42]
Chain-of-Thought Prompting	Use multi-step reasoning prompts, forcing the model to explain intermediate steps.	"Think step-by-step" style prompts.	Can reduce blind guessing by decomposing logic ([39] blog.biostrand.ai) ([24] blog.biostrand.ai). May not eliminate errors and makes prompts more complex.	[17†L94-L102] [55†L59-L62]
Model Fine-Tuning	Further train the LLM on high-quality, domain-specific corpora.	Fine-tuning on ChEMBL or DrugBank text.	Can embed correct domain knowledge and reduce obvious mistakes. Computationally expensive; risk of overfitting biases if data is limited ([11] blog.biostrand.ai).	[17†L110-L119]
Knowledge Graphs / Ontologies	Use structured biomedical graphs to validate relationships implicit in outputs.	Integrating BioKG or custom pharma ontology.	Ensures consistency (e.g. known drug-target interactions); flags nonsensical links. Building/updating graphs is labor-intensive.	[17†L31-L34] [49†L38-L42]
Fact-Checking LLM Outputs	Decompose LLM answers into claims and verify each against evidence sources.	HalluMeasure; claim-check models (www.amazon.science).	Provides automated detection of unsupported claims. (www.amazon.science) – can catch complex hallucinations; relies on retrieval quality and domain-specific classifiers.	[36†L44-L49] [35†L39-L44]
Cross-Model Consistency Checks	Compare outputs from multiple LLMs or ensemble methods to flag inconsistencies.	Generate same answer via GPT-4 and Flan-T5.	If answers differ, could indicate uncertainty. However, all models may hallucinate similarly on novel queries.	[3†L59-L64] [19†L202-L210]
Statistical Confidence Scoring	Use the model's internal token probabilities to gauge uncertainty and flag low-confidence answers.	Interpret logits or use calibration techniques.	Could filter out uncertain outputs. LLM "confidence" often poorly calibrated and not reliable enough for safety-critical use.	[17†L59-L66]
Controlled Decoding	Apply constrained decoding methods (e.g. beam search with penalties) to prefer factual consistency.	Factuality-enhanced decoding methods.	Biases generation towards known facts. Effective but may reduce creative diversity.	[28†L1781-L1787]
Prompt Library and Templates	Use structured prompts crafted to minimize hallucination (e.g. zero-shot vs few-shot engineering).	Chain-of-Verification prompts ([39] blog.biostrand.ai).	Good prompts can reduce errors ([10] www.drugdiscoveryonline.com) but rely on prompt design skill; not foolproof.	[3†L67-L75] [17†L94-L102]
Human-in-the-Loop Review	Always have domain experts review and approve AI-generated outputs before action.	Team reviews AI-suggested targets or texts.	Essential for high-stakes decisions ([12] www.drugdiscoveryonline.com) ([6] pmc.ncbi.nlm.nih.gov), captures nuanced errors. Time-consuming and hard to scale across all uses.	[3†L77-L85] [41†L271-L279]
Benchmark Testing	Test the LLM on known-challenge examples (like Med-HALT) to measure hallucination tendencies.	Use specialized test sets for drug queries.	Helps quantify risk and calibrate usage. Developing benchmark is itself a research task (Med-HALT, etc.).	[52†L15-L21] (Med-HALT)

Table 2: *Strategies to detect or prevent AI hallucinations in drug discovery.* Each strategy has trade-offs; typically a combination (especially RAG plus human review) is needed. The cited references illustrate these approaches in practice. For instance, BioStrand recommends RAG and chain-of-thought prompting ([39] blog.biostrand.ai) ([7] pmc.ncbi.nlm.nih.gov), while Amazon's HalluMeasure specifically automates the *fact-checking* of each claim (www.amazon.science).

Retrieval-Augmented and Knowledge-Grounded Generation

One of the most effective approaches is to **ground AI outputs in external source data**. Rather than rely solely on the LLM's implicit memory, retrieval-augmented generation (RAG) fetches relevant documents or database entries to populate the response. In pharma, this could mean having the LLM cite DrugBank for drug properties or clinical databases for disease associations. FDA-approved frameworks and industry platforms increasingly emphasize this: Mendel AI's "Hypercube" uses LLMs only as a front-end, always linking answers back to its curated database of clinical records ([40] www.pharmavoice.com). Sinequa notes that in regulated environments, every AI-generated statement must be "grounded… in verified, cited, auditable data" ([8] www.sinequa.com). Similarly, researchers have integrated specialized LLMs (e.g., ChemCrow, Galactica) with large biomedical knowledge graphs so that molecule suggestions are checked against known chemistry genetics relationships ([41] blog.biostrand.ai) ([16] pmc.ncbi.nlm.nih.gov).

RAG has proven merit. When each LLM response is explicitly anchored to source text (via citations or snippets), unsupported claims become easier to spot. In Amazon's HalluMeasure work, they rely on retrieving context to fact-check claims (www.amazon.science). Fundamentally, RAG shifts the task from "memorize everything" to "retrieve and synthesize," which greatly cuts hallucination. However, success depends on a high-quality corpus and retrieval system. If the knowledge base is incomplete, or the retrieval query fails, hallucinations can still occur.

Prompt Engineering and Reasoning Chains

Prompt design is another line of defense. By carefully phrasing queries and encouraging multi-step reasoning, users can sometimes reduce hallucinations. Techniques like **chain-of-thought (CoT) prompting** coax the model to list intermediate reasoning steps, which often improves factual accuracy. For example, instructing the model to "explain step by step how you arrived at that answer" can catch errors early ([39] blog.biostrand.ai). An extended concept, *Chain-of-Verification*, asks the LLM to verify its own answer(s) afterward ([39] blog.biostrand.ai). These methods exploit the observation that well-structured prompts can mitigate, though not eliminate, blind guessing.

However, prompt fixes have limits. They rely on the model actually "caring" about accuracy versus linguistic flair. In many experiments, even clever prompts failed to fully eliminate hallucinations ([24] blog.biostrand.ai). Prompting is useful but not sufficient in critical pipelines – especially if prompts become very long or complicated.

Model Tuning and Safety Mechanisms

Developers are also enhancing models themselves. Fine-tuning an LLM on curated drug discovery datasets (targeted biomedical corpora, reaction databases, etc.) can reduce hallucinations by aligning its knowledge. Parameter-efficient fine-tuning (LoRA, adapters) allows incorporating new data without retraining the whole model ([11] blog.biostrand.ai). Domain-specific LLMs (like Galactica, Med-PaLM) are built with more medical literature in training. Additionally, reinforcement learning from human feedback (RLHF) is often used to discourage falsehoods.

Decoding algorithms are evolving too. Researchers have devised "factuality-enhanced decoding" that penalizes improbable tokens or cross-checks consistency ([42] pmc.ncbi.nlm.nih.gov). For instance, some methods dynamically consult a knowledge graph during generation to prune unsupported branches. Others integrate a secondary model that flags inconsistency in real-time. These are active research fronts; for now, they supplement (but do not replace) data grounding and human review.

Claim-Level Verification and Automated Checkers

Beyond generating better outputs, one can climb "post-statistical pyramid" and automatically **detect** hallucinations. Amazon's HalluMeasure is an example: it decomposes an AI answer into individual "claims" and verifies each against a reference context (www.amazon.science). Each claim is classified as supported, contradicted, or unknown given retrieved literature. A high rate of unsupported claims signals hallucination. In their EMNLP23 paper, HalluMeasure achieved fine-grained analysis of hallucination types, offering a "hallucination score" for entire outputs. This method could be deployed in drug pipelines: after an LLM suggests a potential target or mechanism, a tool like HalluMeasure would parse each sentence and check against published data and molecular databases to flag dubious assertions.

Another approach is adversarial probing: designing test questions meant to "trap" hallucinations. For example, standardized benchmarks (like Med-HALT in healthcare ([43] aclanthology.org)) present LLMs with trick scenarios where a hallucination would be obvious. Running an LLM through such tests quantifies its trustworthiness in sensitive domains.

Human-in-the-Loop and Governance

Despite all technical fixes, **human oversight remains essential**. The consensus is clear: AI must operate as an aid, not an autonomous decision-maker. DrugDiscoveryOnline emphasizes "the human element" – techniques like Chain-of-

Thought help, but a human must be ready to verify outputs ([12] www.drugdiscoveryonline.com). Indeed, in practical projects researchers pair LLMs with experts who vet every suggestion. In the aforementioned anticocaine addiction study, the team explicitly cross-checked ChatGPT's proposals by (1) literature and (2) expert reasoning, accepting only well-supported insights ([6] pmc.ncbi.nlm.nih.gov). Similarly, regulatory submissions currently require human signoff; any AI-used component must be reviewed for accuracy.

Organizations are implementing governance frameworks to enforce this. For example, Elsevier's Responsible AI principles (for research tools) include "traceability" (logging all AI prompts/outputs) and "verification"—requiring that claims be linked to source material. In practice, drug companies often treat hallucinations as "system crashes": they monitor output likelihood and have procedures to discard or redo hallucinated cases ([37] www.drugdiscoveryonline.com).

Data Analysis: Quantifying Hallucinations

Several studies have attempted to measure hallucination prevalence and impact quantitatively, providing insight into the scope of the problem:

- **Citation Fabrication:** The stem-cell essay study found ~~24.4% of references from ChatGPT were faulty (15.1% fabricated + 9.3% erroneous) ([29] pmc.ncbi.nlm.nih.gov). Another analysis of 50 medical abstracts led by Gao et al. had only 36.2% accurate references, implying ~63.8% fabricated ([44] pmc.ncbi.nlm.nih.gov). These numbers are alarmingly high, showing naive ChatGPT usage for literature can be wildly unreliable.
- **Model Survey:** In one broad test of 11 LLMs on various tasks, the best model still hallucinated ~3% of the time (worst was 27%) ([26] blog.biostrand.ai). This suggests even top-tier AIs are not infallible. The authors caution that in drug discovery, even 3% can derail projects ([28] blog.biostrand.ai).
- **Task-specific:** In chemistry education, only 28% of ChatGPT's mechanism explanations were entirely correct ([36] pubs.acs.org). This low success rate on a standard chemistry task underscores how many "plausible" answers contain holes.
- **Return on Distortion:** Morgan Stanley quantified the upside of improved accuracy: just boosting early-stage success rates slightly could yield 50+ new drugs in a decade ([17] www.drugdiscoveryonline.com). By implication, hallucinations that *decrease* success rates by a few percentage points (through wrong leads or hypotheses) can have equally large negative cost.

These data points, drawn from literature, confirm that hallucinations are not merely an anecdotal annoyance but a measurable, systemic risk in applying AI to biomedical domains. They also underscore how any detection strategy must handle non-trivial frequencies of error.

Case Study Analysis

We briefly examine how specific organizations and projects have dealt with hallucinations:

- **Mendel AI (PharmaVoice)**: Mendel's Hypercube platform combines an LLM with a logic-based engine. It builds an internal index of patient records, guaranteeing each AI statement can be traced back to actual data ([40] www.pharmavoice.com). In effect, it forbids "free-form" generation by forcing AI outputs into a database query framework. Philosophically, this mimics a search engine: AI predictions are only allowed to rerank or summarize retrieved facts. Early results in customer trials suggest this approach dramatically reduces hallucination risk and re-establishes trust ([40] www.pharmavoice.com).
- **In-house RAG Pipelines:** Many pharma companies are creating their own RAG systems. One approach, exemplified by BioStrand's LENS platform, uses a dynamic knowledge graph of billions of bio-entities ([45] blog.biostrand.ai). Any LLM suggestion is cross-validated against this graph. For instance, if the LLM mentions a drug-

target link, LENS checks it against known sequences/functions. This hybrid AI (LLM + symbol-based checks) is currently used in prototypes to filter AI suggestions before human scientists see them.

- **OpenAI & Academic Efforts:** Following the Cureus and other papers, AI practitioners widely acknowledge hallucinations. OpenAI and other labs are actively researching mitigation. For example, GPT-4 introduced a web browsing plugin (May 2023) to fetch current data, reducing hallucination from its older knowledge cutoff ([46] pmc.ncbi.nlm.nih.gov). Academic workshops (e.g., Duke/FDA AI workshop 2022 ([47] www.fda.gov)) and consortia have been convened to develop standards for AI credibility. These collaborative efforts indicate that catching hallucinations is a shared priority in the field.

Future Directions and Recommendations

As AI becomes pervasive in drug discovery, addressing hallucinations will remain a moving target requiring continual vigilance. Based on current trends and expert opinion, the following directions are crucial:

1. **Standardized Benchmarks**: The community needs more domain-specific hallucination benchmarks (like Med-HALT in healthcare) to test models on realistic drug-centric tasks. New challenges should simulate common pharma queries (e.g. ask for dosing of a rare drug) to expose weaknesses. Models can then be iteratively improved using these tests.

2. **Explainable AI and Transparency**: Future LLMs might be built with explainability in mind. Rather than black-box text, they could present structured reasoning or provenance (e.g. “This recommendation came from X sources”) to allow easier vetting. This would help scientists quickly judge reliability.

3. **Regulatory and Industry Guidelines**: We expect official standards. The FDA’s draft guidance ([13] www.fda.gov) is a first step; likely final rules will explicitly require risk assessments of AI hallucination (e.g. how much of an output is AI-generated vs verified). Industry bodies (like DIA or industry consortia) may publish whitepapers on best practices, akin to Good Machine Learning Practices (GMLP) but specific to generative AI.

4. **Hybrid Human-AI Workflows**: Successful deployments will build systems where AI and humans collaborate. Routine or low-risk tasks may be automated, but any critical decisions or novel outputs will trigger human review flags. User interfaces will likely incorporate disclaimers and calls for confirmation, as in medication prescribing software.

5. **Technical Advances**: On the research side, solutions are evolving. Multimodal models could check consistency across text, structure, and data (e.g. an LLM’s proposed molecule could be immediately run through a chemistry validator). Techniques like self-reflection (LLMs tasking themselves with error-checking) and embodiment in tools (where the LLM operates through APIs with constraints) show promise. The EMPOWER framework (2025) and other systematic processes aim to iteratively refine LLM responses, cutting hallucinations significantly ([48] relixir.ai).

6. **Education and Culture**: Perhaps most importantly, data scientists and drug researchers must internalize that AI outputs **require verification**. As the Cureus editorial urges, outputs should never be blindly trusted ([49] pmc.ncbi.nlm.nih.gov) ([50] pmc.ncbi.nlm.nih.gov). Training programs for chemists and biologists increasingly include “AI literacy” modules. The culture is shifting from “if ChatGPT says it, it’s true” to “let’s double-check what the model says.” This cultural change is already underway in academic publishing and will grow in industry.

Conclusion

AI hallucinations in drug discovery represent a serious challenge but not an insurmountable one. Real-world evidence shows that leading LLMs and AI models **do** produce plausible but false information across the pipeline – from invented scientific citations to bogus medical claims ([1] pmc.ncbi.nlm.nih.gov) (www.amazon.science). These errors can misdirect research, waste resources, and even threaten patient safety ([2] www.sinequa.com) ([3] blog.biostrand.ai). However, by combining multiple strategies – rigorous grounding in vetted data, careful prompt engineering, post-hoc fact-checking, and human expert oversight – the risk can be managed. Companies and regulators are already moving to embed such checks: FDA guidelines, internal QA processes, and new AI tools (like HalluMeasure (www.amazon.science)) are being adopted to “catch” hallucinations.

Looking forward, we anticipate a dual trajectory: on one hand, ever-more-sophisticated AI generation will continue to surprise and occasionally mislead. On the other, growing infrastructure around these models will provide “safety nets.” By emphasizing transparency, auditability, and accountability in AI-driven drug discovery, the field can harness the transformative potential of AI while minimizing its blind spots. The future likely will see AI as a collaborative partner – one that is constrained by human expertise and rigorous validation at every step. In this way, the “creative missteps” of AI ([14] kingy.ai) can be allowed only where they spur innovation, not where they introduce unsafe errors. As one commentator puts it, AI hallucinations must be treated as part of the design: “sometimes problems can be accepted as part of the solution,” akin to acceptable side-effects in pharmacology ([51] www.drugdiscoveryonline.com). But just as side-effects are carefully monitored, so too must AI’s hallucinations be monitored and corrected.

By continually integrating evidence-based detection methods, updating regulatory frameworks, and fostering an informed user base, the drug discovery community can navigate the “dark side” of AI to its advantage. In doing so, we can ensure that AI remains a powerful ally in developing new therapies, not an unchecked source of misinformation.

External Sources

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC9939079/#:~:We%20...

[2] https://www.sinequa.com/resources/blog/generative-ai-a-new-frontier-in-pharmaceutical-drug-development-and-clinical-trial-analysis/#:~:A%20r...

[3] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:This%...

[4] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:How%2...

[5] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:1,pro...

[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC11021135/#:~:It%20...

[7] https://pmc.ncbi.nlm.nih.gov/articles/PMC12546459/#:~:these...

[8] https://www.sinequa.com/resources/blog/generative-ai-a-new-frontier-in-pharmaceutical-drug-development-and-clinical-trial-analysis/#:~:enter...

[9] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:There...

[10] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:Techn...

[11] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:Fine...

[12] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:and%2...

[13] https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions#:~:A%20k...

[14] https://kingy.ai/news/blog-post-title-ai-hallucinations-as-a-catalyst-for-faster-drug-discovery/#:~:Artif...

[15] https://pmc.ncbi.nlm.nih.gov/articles/PMC12546459/#:~:devel...

[16] https://pmc.ncbi.nlm.nih.gov/articles/PMC12546459/#:~:diagn...

[17] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:It%27...

[18] https://pmc.ncbi.nlm.nih.gov/articles/PMC8132984/#:~:%E2%8...

[19] https://pmc.ncbi.nlm.nih.gov/articles/PMC8132984/#:~:Intro...

[20] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:AI%20...

[21] https://www.pharmavoice.com/news/mendel-ai-hallucination-hypercube-google-cloud/725290/#:~:Hallu...

[22] https://www.pharmavoice.com/news/mendel-ai-hallucination-hypercube-google-cloud/725290/#:~:%E2%8...

[23] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:some%...

[24] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:At%20...

[25] https://www.linkedin.com/posts/ericneuman_this-is-a-game-changing-discovery-hallucinations-activity-7370745326598729728-JukA#:~:Eric%...

[26] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:of%20...

[27] https://www.linkedin.com/pulse/150-billion-hallucination-problem-why-expert-data-silent-tony-medrano-ndydc#:~:Follo...

[28] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:gener...

[29] https://pmc.ncbi.nlm.nih.gov/articles/PMC10553015/#:~:Stem%...

[30] https://pmc.ncbi.nlm.nih.gov/articles/PMC8132984/#:~:As%20...

[31] https://pmc.ncbi.nlm.nih.gov/articles/PMC8132984/#:~:Figur...

[32] https://www.pharmavoice.com/news/mendel-ai-hallucination-hypercube-google-cloud/725290/#:~:%E2%8...

[33] https://pmc.ncbi.nlm.nih.gov/articles/PMC10553015/#:~:journ...

[34] https://pmc.ncbi.nlm.nih.gov/articles/PMC10553015/#:~:match...

[35] https://pmc.ncbi.nlm.nih.gov/articles/PMC9939079/#:~:We%20...

[36] https://pubs.acs.org/doi/10.1021/acs.jchemed.4c00235#:~:While...

[37] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:Summa...

[38] https://www.pharmavoice.com/news/mendel-ai-hallucination-hypercube-google-cloud/725290/#:~:Hyper...

[39] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:based...

[40] https://www.pharmavoice.com/news/mendel-ai-hallucination-hypercube-google-cloud/725290/#:~:Simil...

[41] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:When%...

[42] https://pmc.ncbi.nlm.nih.gov/articles/PMC12546459/#:~:accur...

[43] https://aclanthology.org/2023.conll-1.21/#:~:Med,H...

[44] https://pmc.ncbi.nlm.nih.gov/articles/PMC10553015/#:~:and%2...

[45] https://blog.biostrand.ai/mitigating-llm-hallucinations#:~:Holis...

[46] https://pmc.ncbi.nlm.nih.gov/articles/PMC11021135/#:~:Moreo...

[47] https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions#:~:In%20...

[48] https://relixir.ai/blog/cutting-hallucinations-25-percent-empower-framework-gpt-4o-medical-prompts#:~:Cutti...

[49] https://pmc.ncbi.nlm.nih.gov/articles/PMC9939079/#:~:early...

[50] https://pmc.ncbi.nlm.nih.gov/articles/PMC9939079/#:~:The%2...

[51] https://www.drugdiscoveryonline.com/doc/beware-ai-hallucinations-0001#:~:condi...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.