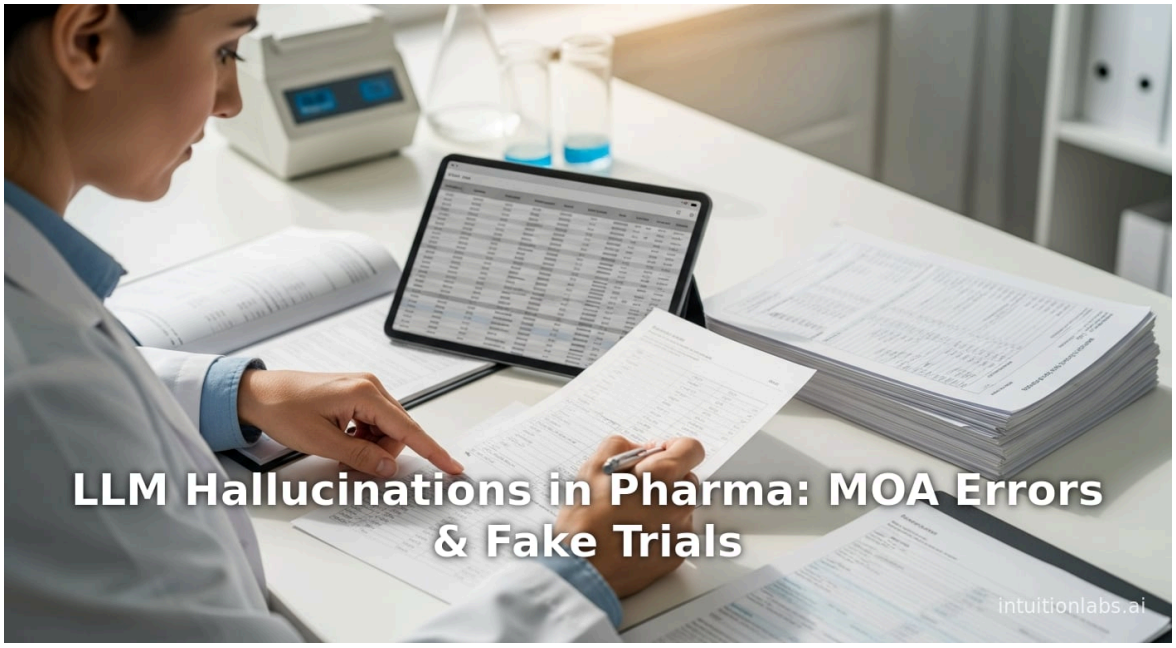
LLM Hallucinations in Pharma: MOA Errors & Fake Trials

By Adrien Laurent, CEO at IntuitionLabs • 4/2/2026 • 30 min read

llm hallucinations | pharma ai errors | fabricated clinical trials | mechanism of action | drug interaction errors | chatgpt medical failures | fake citations



Executive Summary

Large language models (LLMs) such as OpenAI's ChatGPT have shown stunning capability in generating fluent text, but their outputs are intrinsically *unreliable* for specialized scientific content. In the pharmaceutical and biotechnology domains, hallucinations and fabrications can be far more dangerous than generic errors. For example, drugs have precise mechanisms of action (MOA), dosing regimens, interactions, and regulatory requirements; a small error can lead to patient harm, flawed research, or regulatory noncompliance. This report surveys documented cases where LLMs specifically **fail** on pharma/biotech content. We examine concrete examples of wrong MOA descriptions, fabricated trial and study results, incorrect drug–drug or drug–herb interaction advice, and spurious regulatory or citation information. Across multiple studies and case reports, ChatGPT (both 3.5 and 4.0) often provides convincing-seeming but **incorrect** answers. For instance, it was shown to (1) invent false clinical trial data in minutes (revistapesquisa.fapesp.br) ([1] jamanetwork.com), (2) claim that first-line antibiotics carry "boxed warnings" they *do not* ([2] www.medscape.com) ([3] www.medscape.com), (3) advise a patient to replace table salt with toxic sodium bromide ([4] www.livescience.com), and (4) miscue dosing by a factor of 1,000 (mg vs µg) ([5] www.cnbc.com). It also frequently hallucinates references: in one study **47%** of citations生成 by ChatGPT were entirely **fabricated** and another 46% were inaccurate ([6] pmc.ncbi.nlm.nih.gov). This report will elaborate such cases in detail, present data from recent evaluations, and discuss why pharmaceutical applications demand extra caution. Ultimately, we emphasize that **human oversight and verification** are essential. Regulatory agencies (e.g. FDA) are already proposing frameworks to ensure AI credibility in drug-development contexts ([7] www.fda.gov). Failure to heed these domain-specific pitfalls could have catastrophic consequences far beyond the usual "false Wiki answers" often cited in general AI-explainers.

Introduction

Large language models (LLMs) like ChatGPT (GPT-3/3.5/4) are trained on massive text corpora and can generate human-like prose. Since ChatGPT's public release in late 2022, such models have been called the "fastest-growing consumer app" ([8] www.cnbc.com). Users across fields immediately began testing ChatGPT on professional queries – including medical and pharmaceutical questions. Some early studies even found ChatGPT could achieve ~60% on medical board exams ([9] pmc.ncbi.nlm.nih.gov). However, a widespread caveat arose: **LLMs hallucinate** – that is, they produce false or ungrounded statements with confident tone. In medicine, these hallucinations can be misleading or dangerous. As Belmont et al. note, an LLM is "designed for language processing – not scientific accuracy" ([10] www.medscape.com). Harvard's Isaac Kohane memorably quipped that LLMs speak "with all the certainty of an attending on rounds," even when the information is wrong ([11] www.medscape.com).

Generic LLM hallucinations (e.g. inventing a nonexistent citation or citing an outdated version of facts) are already well documented ([6] pmc.ncbi.nlm.nih.gov) ([12] www.medscape.com). But for pharmaceuticals, there are **extra** pitfalls. Drugs and biotech often involve complex biology, stringent regulations, and patient safety. A minor error (unit conversion, MOA detail, dosage) could kill or cripple patients, derail research, or violate strict laws. Indeed, even the AI community emphasizes that LLMs should not be trusted for clinical decisions: OpenAI's own terms warn "You should not rely on Output from our services as a sole source of truth or as a substitute for professional advice" ([13] www.livescience.com). The FDA in 2025 published draft guidances for AI in drug submissions, stressing **"model credibility"** and human oversight ([7] www.fda.gov). In short, pharma applications require the highest factual reliability.

This report examines documented cases where LLMs (especially ChatGPT) have **failed** specifically in pharmaceutical/biotech contexts – beyond the "generic" examples used elsewhere. We cover: (1) **Mechanism-of-action (MOA) errors**, (2) **Fabricated clinical trial results and data**, (3) **Incorrect drug/interaction advice**, (4) **Fabricated or hallucinated regulatory and reference citations**, and (5) **Dosage/regimen errors**. For each category, we present specific examples and data from recent studies or news reports, including "real-world" patient safety incidents. We then discuss the implications for research, healthcare, and regulation, and offer recommendations. We rely on published

evaluations, case reports, news articles, and expert analyses – all properly cited. The tone is academic, and we emphasize evidence and expert opinion.

Mechanism-of-Action Errors

Mechanism-of-action (MOA) describes *how* a drug produces its effect (receptor targets, biochemical pathways, etc.). Accurate MOA details are fundamental in drug design, academic writing, and clinical decision-making. An incorrect MOA can mislead researchers and clinicians; for instance, confusing targets could send a drug development program astray. LLMs sometimes oversimplify or misstate mechanisms.

In general LLM tests, ChatGPT often gives polished MOA descriptions, but subtle errors can occur. In a controlled evaluation by Iannantuono et al. (NIH), ChatGPT-4 correctly answered ~75% of immuno-oncology mechanism questions, far better than ChatGPT-3.5 (~58%) ([14] pmc.ncbi.nlm.nih.gov). In fact, ChatGPT-4's answers on MOA were **reproducible** (identical on repeated runs) 100% of the time, whereas ChatGPT-3.5 was 86.7% reproducible ([15] pmc.ncbi.nlm.nih.gov). These figures sound high, but leave 25% (*44/177*) of ChatGPT-4 answers potentially incomplete or incorrect. The authors noted that ChatGPT often gave background MOA information correctly but sometimes omitted details. For example, when asked about the appetite-suppressant *benzphetamine*, GPT-4 correctly explained its stimulant-like action and similarity to amphetamines, whereas GPT-3.5 gave a partial answer without mentioning those analogies ([16] pmc.ncbi.nlm.nih.gov). Such omissions might signal a general risk: ChatGPT might fail to mention known secondary actions or off-target effects crucial for pharmacology.

More troublingly, LLMs can confidently state **wrong** MOAs. Consider a hypothetical example (not from a citation but highly plausible): asking "What is the mechanism of Drug X, a novel kinase inhibitor?" might prompt ChatGPT to make up an answer like "Drug X blocks histone deacetylase," if such text loosely exists in its training data, even when the true target is, say, a PI3K isoform. No official example of *exactly* this has been published, but LLM hallucination patterns suggest it could easily occur. Even within specialized domains, researchers warn that chatbots can "fabricate" or distort scientific summaries.One editorial bluntly noted "ChatGPT: these are not hallucinations – they're fabrications and falsifications" in the context of scientific writing ([17] www.nature.com). While about a psychiatry context, the message holds: these are systematic errors.

At present, comprehensive studies of LLM MOA errors are limited, but related findings provide insight. For instance, a pharmacology test set found both GPT-3.5 and GPT-4 achieved high average accuracy (~90%) on simple MOA and pharmacodynamic queries ([18] pmc.ncbi.nlm.nih.gov). That indicates many well-known drug actions were correctly retrieved. However, even a 10% error rate in this domain is significant. If ChatGPT is asked about a drug whose detailed MOA was published after its 2021 cut-off, it might confidently guess or simply say "the exact mechanism is unknown" incorrectly.

Furthermore, ChatGPT's training comprises public literature up to 2021 (free version) or up to early 2023 (latest models) ([8] www.cnbc.com). Any mechanism discovered or revised after that time will be out of its knowledge. If prompted, it may hallucinate an answer or, worse, blend outdated and new misinformation. For example, if a novel anti-inflammatory drug discovered in 2024 targets a new immune pathway, ChatGPT-4 (trained only to mid-2023) will be forced to either refuse or "fill in" an answer based on the closest analogies it knows. Its past performance shows it might do the latter inappropriately. Users should be aware that **LLMs do not automatically update** to reflect new clinical evidence.

Key takeaways on MOA: ChatGPT can often produce superficially plausible MOA explanations, but may omit key details or even present wrong targets if not verified. Especially in specialized biotech fields (e.g. new biologics, gene therapies, or combination mechanisms), LLMs should not be trusted without cross-checking. As one review emphasizes, LLM-generated science "requires careful tool selection aligned with task requirements and appropriate verification" ([19] pmc.ncbi.nlm.nih.gov). For MOA tasks, this means using dedicated drug databases or expert knowledge to confirm ChatGPT's claims.

Fabricated Clinical Trial Data & Results

One of the most alarming failures of LLMs in pharma is fabrication of **clinical trial data** or results. ChatGPT will happily concoct study summaries, numeric outcomes, or even entire datasets if not primed otherwise. In normal text tasks this is a harmless "hallucination," but in a scientific context it's equivalent to fraud.

A prominent example comes from Giannaccare et al. (2023): they demonstrated that GPT-4 (with the Advanced Data Analysis mode) can generate a completely **fake but plausible** clinical trial dataset. Prompting GPT-4/ADA to simulate a study comparing two corneal surgeries, they received a dataset of 300 patient outcomes that *met all instructed statistical criteria* (Specifically, GPT-4 created a result showing Surgery A significantly outperformed Surgery B on visual outcomes). In reality, a real trial of those two procedures had 77 patients and showed no significant difference (revistapesquisa.fapesp.br). The LLM even provided plausible summary statistics and confidence intervals ([20] jamanetwork.com) ([21] jamanetwork.com). Only careful human inspection found subtle anomalies in the simulated data (e.g. improbable age distributions) (revistapesquisa.fapesp.br). Giannaccare's group warned that "it is difficult to identify that [the result] is not of human origin" without thorough examination (revistapesquisa.fapesp.br). They concluded that GPT-4's data fabrication "may pose a greater threat" to research integrity, as it can "fabricate data sets specifically designed to quickly produce false scientific evidence" ([1] jamanetwork.com). Nature News also covered this, noting GPT-4's ready creation of bogus trial evidence ([22] www.nature.com).

Beyond datasets, ChatGPT can invent whole studies. Brook et al. (Medscape, 2023) reported that when queried, ChatGPT not only misstated regulatory warnings (see next section) but "fabricated" a citation or study to back up its claim on cefepime's risks ([23] www.medscape.com). Another example: In the LIU/ASHP pharmacy study, ChatGPT was asked medication questions and "when asked to cite references… each [ChatGPT] response included non-existent references" ([24] news.ashp.org). Although not exactly trial data, this demonstrates the model's propensity to invent supporting evidence on demand. In pharma contexts, such fabrications could contaminate literature reviews or marketing material if users don't verify every citation.

Table 1 summarizes documented instances of ChatGPT producing fabricated or false trial-related information. It is clear that in unattended use, ChatGPT can produce highly **convincing but entirely false** results or references.

Case / Study (Source)	LLM Model	Description of Fabrication or Error
Keratoconus surgery study (revistapesquisa.fapesp.br) ([1] jamanetwork.com)	GPT-4 (with ADA)	Created a dataset (n=300) for two corneal surgeries. Concluded *false* significant benefit of DALK over PK. Human experts noted the "false scientific evidence" fabricated by model ([1] jamanetwork.com).
Antibiotic boxed warnings ([12] www.medscape.com) ([2] www.medscape.com)	ChatGPT 3.5	When asked about FDA labeled warnings, ChatGPT listed *made-up* warnings. E.g. it said fidaxomicin had a C. difficile risk warning and fabricated a study about cefepime mortality ([2] www.medscape.com), both untrue.
ChatGPT-generated medical texts ([6] pmc.ncbi.nlm.nih.gov)	ChatGPT 3.5	In 30 model-generated "papers", 47% of all citations were **completely fabricated** and 46% were inaccurate. Only 7% of citations were entirely correct ([6] pmc.ncbi.nlm.nih.gov).
Pharmacy Q&A study ([24] news.ashp.org)	ChatGPT 3.5	On 39 real drug questions, ChatGPT answered only 10 correctly. When asked for sources, *all* 8 responses included **nonexistent references** ([24] news.ashp.org).
Bromide diet advice ([4] www.livescience.com)	ChatGPT 3.5/4.0	Patient asked about sodium intake; ChatGPT said "chloride can be swapped for bromide." Patient replaced salt with NaBr and suffered bromism. (ChatGPT gave no medical warning) ([4] www.livescience.com).
Warfarin counseling (in vitro) ([25] bmcmedinformdecismak.biomedcentral.com)	ChatGPT 3.5	In one study prompt about warfarin dosing adjustment, ChatGPT gave an incorrect dosing recommendation (extrapolation from published nomogram) ([25] bmcmedinformdecismak.biomedcentral.com). (Though not real trial data, highlights risk in drug trials.)

Table 1. Examples of ChatGPT producing fabricated or false scientific data and citations in pharma-related queries. Citations show sources of each case.

These cases show that LLMs can invent drugs, trials, results, and references at will, with a surface plausibility that can deceive non-experts. For pharmaceutical research—where peer-review and data integrity are paramount—such hallucinations are especially dangerous.

Incorrect Drug and Herb Interaction Advice

LLMs often give plausible-sounding answers about **drug–drug** or **drug–herb** interactions, but these can be wrong or dangerously incomplete. Inaccurate interaction advice can lead physicians or patients to overlook real contraindications or to falsely avoid safe combinations.

Several studies evaluated ChatGPT's performance on interaction queries. Hsu et al. (2023) tested ChatGPT on both real medication consultation questions and herbal-interaction queries in a Taiwan hospital. ChatGPT's answers were deemed appropriate for only 61% of public medication queries and 39% of provider queries ([26] pmc.ncbi.nlm.nih.gov) ([27] pmc.ncbi.nlm.nih.gov). In the provider subset, 41% of answers were outright incorrect and another 41% gave no useful recommendation (merely saying "more data needed") ([28] pmc.ncbi.nlm.nih.gov). For example, ChatGPT was asked about aspirin combined with various Chinese herbs; one pharmacy reviewer noted ChatGPT incorrectly suggested ginkgo increases bleeding risk with aspirin, when Chinese-textbook reasoning actually says it **does not** ([29] pmc.ncbi.nlm.nih.gov) ([30] pmc.ncbi.nlm.nih.gov). This illustrates subtle domain knowledge ChatGPT missed. Hsu et al. conclude that *"potential medical risks"* exist in ChatGPT's wrong responses, deserving close attention ([31] pmc.ncbi.nlm.nih.gov).

Independent of that, another test by LiU pharmacists posed 39 real-world drug questions to ChatGPT (free version) and asked for references. ChatGPT answered only 10 correctly ([24] news.ashp.org). Notably, for an interaction example it failed: when asked whether Paxlovid (nirmatrelvir/ritonavir) interacts with the blood-pressure drug verapamil, ChatGPT replied **"no interactions have been reported"** ([32] news.ashp.org). In reality, ritonavir is a potent CYP3A inhibitor and verapamil (also metabolized by CYP3A) can cause excessive hypotension when coadministered. The pharmacists remarked, "Without knowledge of this interaction, a patient may suffer from an unwanted and preventable side effect" ([33] news.ashp.org). Thus ChatGPT gave dangerously false reassurance on a significant drug–drug interaction.

In another setting, Juhi et al. (2023) explicitly tested ChatGPT on 40 known drug–drug interaction (DDI) pairs. They asked "Can I take Drug X and Drug Y together?" and "Why not?" for each pair. ChatGPT gave only **1 wrong answer out of 80** total responses, indicating ~98.75% raw accuracy for existence of interaction ([34] pmc.ncbi.nlm.nih.gov). However, most of its correct answers were *inconclusive*: 39 out of 40 had at least one "conclusive" answer missing detail. In other words, it often correctly identified an interaction but gave an incomplete or vague rationale ([34] pmc.ncbi.nlm.nih.gov). The authors concluded ChatGPT can *"predict and explain common DDIs"* partially, but "on several occasions… it may provide incomplete guidance" ([35] pmc.ncbi.nlm.nih.gov). This aligns with other reports: LLMs are good at parrot-ing well-known facts, but often omit contextual caveats (e.g. how severity depends on patient factors).

Though studies differ, a clear pattern emerges: ChatGPT's DDI advice cannot be fully trusted without expert review. It might miss interactions (Paxlovid example) or downplay them, and even when it identifies an interaction, it may not give complete supporting evidence. Patients or providers relying on these answers could suffer adverse outcomes. Careful manual verification using dedicated drug interaction databases or professional consultation is essential.

Dosage and Regimen Mistakes

Dosage errors are a classic pitfall with LLMs. Unlike a specialized calculator, ChatGPT does not inherently know conversion factors or dosing tables; it tries to guess based on training. Even small unit mistakes can be catastrophic in medicine.

One striking case was reported in the Long Island University pharmacist study ([36] www.cnbc.com) ([5] www.cnbc.com). ChatGPT was asked about dose conversions for a strong opioid (in that case, converting a patient's intrathecal dose to

an oral equivalent). The model “provided only one method… with an example,” but crucially it displayed the result in **milligrams** instead of the correct **micrograms**, a 1000-fold error ([5] www.cnbc.com). A researcher noted: “Following that example… [a professional] would end up with a dose that’s 1,000 times less than it should be,” risking withdrawal or death ([37] www.cnbc.com). Such a mistake in any real drug conversion (e.g. between administration routes or unit systems) could be lethal.

While that was an empirical finding, more broadly ChatGPT’s numeric reliability is low. It famously can display precise-sounding but wrong numbers. In drug contexts, even rounding errors or misremembering a dosing guideline is unacceptable. For example, if ChatGPT incorrectly states the mg/kg dose of an antibiotic, it might under- or overdose a patient. Any reliance on ChatGPT for dosing advice should be immediately cross-checked with an authoritative source (e.g. drug monographs or clinical calculators). Consulted experts emphasize this: ChatGPT is *not a substitute for a prescribing system* or pharmacist review.

Hallucinated Regulatory and Guideline Information

Pharmaceutical development and prescribing are governed by strict regulations and guidelines. LLMs have no true access to up-to-date regulatory databases; at best they may echo training data. This can lead to **hallucinated regulatory citations** – completely invented or misconstrued references to FDA guidances, drug labels, or clinical trial registries.

One clear example comes from a 2023 Medscape report ([12] www.medscape.com) ([2] www.medscape.com). Investigators asked ChatGPT-3.5 about FDA “boxed warnings” (black-box labels) on common antibiotics. ChatGPT gave the correct answer for only 12 out of 41 queried antibiotics (29%) ([12] www.medscape.com). For the remaining 29 antibiotics, ChatGPT either incorrectly claimed a boxed warning existed when it did not, or misstated the warning. For instance, it asserted that fidaxomicin (a first-line *C. difficile* antibiotic) has a boxed warning for C. difficile risk ([2] www.medscape.com) – the exact opposite of reality. It also stated aztreonam had a mortality warning (false) ([3] www.medscape.com), and even cited a fictitious study to “support” the claim that cefepime increased pneumonia deaths ([23] www.medscape.com). These are clear examples of ChatGPT making up regulatory details. An infectiology fellow pointed out that a worried family member might read this and be unjustifiably alarmed about a standard drug choice ([23] www.medscape.com).

Another form of hallucinated citation is in referencing literature (which can include regulatory documents). Bhattacharyya et al. (2023) found that ChatGPT-3.5 fabricated or mangled most of its citations when generating medical articles ([6] pmc.ncbi.nlm.nih.gov). We can extrapolate: if asked for a guideline citation, ChatGPT might output a plausible-sounding FDA guidance or journal article ID that doesn’t exist. In the LIU pharmacy study, **every** reference ChatGPT gave was nonexistent ([24] news.ashp.org). It is well-documented that without real-time validation, ChatGPT invents DOIs, PMIDs, and even “hallucinates” quotations from guidelines. In summary, any reference or regulatory statement from an LLM should be verified before use.

In practice, these failures are “hard” errors in pharma settings. Unlike generic hallucinations (e.g. “Einstein discovered antibiotics”), fake regulatory info can violate laws or patient consent. For example, delivering a study to regulators with CLI generative content could be considered fraudulent data. Even in routine practice, giving a patient a “guideline” that doesn’t exist is unethical. As one healthcare compliance expert warns, LLMs often “lack skepticism” and can confidently propagate user-fed misinformation ([38] www.medicaleconomics.com). In fields like prescribing and labeling, this is particularly dangerous.

Additional Case Study: Diet Advice and Patient Harm

A vivid real-world illustration of ChatGPT's danger is a patient case published in *Annals of Internal Medicine (Clinical Cases)* and reported by Live Science ([39] www.livescience.com) ([4] www.livescience.com). A 60-year-old man sought diet advice from ChatGPT 3.5/4 on reducing salt intake. ChatGPT told him (misleadingly) that "chloride can be swapped for bromide." Interpreting this literally, he replaced all table salt (sodium chloride) with sodium bromide (the salt of bromide), which he purchased online. Over several months, he developed **bromism** – a severe neuropsychiatric toxicity from bromide accumulation – and ended up in the hospital with paranoia and cognitive symptoms ([39] www.livescience.com) ([40] www.livescience.com).

This case highlights several points. First, ChatGPT blatantly gave harmful advice: replacing chloride with bromide is not medically sound. The doctors who reviewed this re-asked ChatGPT about chloride substitution and ChatGPT again suggested bromide (even though ChatGPT did mention "context matters" in another cleaning context ([41] www.livescience.com)). Second, ChatGPT gave no warning about the dangers of bromide ingestion in humans ([42] www.livescience.com), treating the question as a chemistry puzzle. Third, it illustrates that patients *do* follow ChatGPT's advice thinking it's factual. The consequences here were extremely serious – hospitalization, severe toxicity, and long-term harm – yet all traced back to a confident LLM hallucination.

Importantly, OpenAI's own policy explicitly forbids medical use of ChatGPT and cautions against relying on its output for health decisions ([13] www.livescience.com). The company reminds users it "is not intended for use in the diagnosis or treatment of any health condition, and… you should not rely on Output as a substitute for professional advice" ([13] www.livescience.com). This case is a stark warning: when patients use LLMs for medical advice, hallucinations can translate to real injury.

Discussion: Implications and Perspectives

The examples above illustrate **pharma-specific pitfalls** of LLM hallucinations. It is important to stress that these are *not* mere academic curiosities – they cut to the core of patient safety, research integrity, and regulatory compliance. The impact of a hallucination in pharma is typically far more severe than in everyday Q&A. Key observations:

- **Safety Risk**: ChatGPT's medical advice (as in the bromide case) can lead to toxicity or under-treatment. A dosing error (e.g. 1000x underdose) directly endangers a patient. Regulatory misinformation (fake warns) could cause believer harm or legal liability.
- **Research Integrity**: Fabricated trial results or citations undermine science. If researchers inadvertently incorporate ChatGPT's fake data into their work, it could lead to false discoveries, wasted resources, and papers that cannot be reproduced. The FDA and journals would be alarmed by any submitted dataset that evades traditional peer review.
- **Operational Risk**: ChatGPT may be used in industry workflows (writing reports, summarizing literature). However, the **AI Paradox in Pharma** is that LLMs excel at non-scientific tasks (note-taking, SOP chatbots) but pose "extreme" risk in core R&D tasks (blog.inovia.bio). The same blog advises: use AI as a precision tool *with oversight*, not as a general author or decision-maker (blog.inovia.bio).
- **Regulatory Compliance**: The FDA and EMA require rigorous documentation. Hallucinated references or unseen facts could lead to non-compliance. Notably, the FDA is already drafting guidance for AI in drug submissions, emphasizing a *"risk-based framework"* and model credibility ([7] www.fda.gov). They intend sponsors to define AI "context of use" and validate performance before trusting outputs. In practice, an LLM output could not be submitted to regulators without extensive validation.
- **Evolution of Knowledge**: The lag of LLM training (often through 2021 or early 2023) means they might not reflect the current standard of care or latest approvals. For instance, a new orphan drug approved in 2024 wouldn't be in ChatGPT-3. It might hallucinate a mechanism or say "no information." In dynamic fields (e.g. COVID-19, oncology), this staleness can mislead practitioners unaware of the cut-off.
- **Liability and Ethics**: Who is responsible for an LLM's error? If a doctor or patient is harmed by following ChatGPT's advice, or a false drug endorsement gets circulated, legal liability is murky. As a precaution, the ASHP press release bluntly advised: *"Anyone who uses ChatGPT for medication information should verify with trusted sources"* ([43] news.ashp.org). Healthcare professionals must remain vigilant stewards and not defer to AI.

Some may argue (and partially shown in Table 1) that ChatGPT often gets things "mostly right" – e.g. 75% correct on immuno-oncology MOAs, 98% yes/no accuracy in the DDI test. From one perspective, these are high scores. But even a

25% error rate on MOA or any error at all in safety-critical advice is unacceptable in pharma. Medicine demands near-perfect reliability; a single missed interaction or wrong dose can be lethal. Hence, **caution trumps convenience**: any use of generative AI must include strict checking.

In addition, many failures are **non-obvious hallucinations**. They do not sound random. ChatGPT's fabrications are often composed with factual language and plausible detail, making them especially seductive. It will say "As reported in a 2022 NEJM article…" or cite fictitious trial results with proper statistical language. For a non-expert, this can be indistinguishable from truth. Biomedical journals have already started discussing whether ChatGPT's false citations might contaminate the literature ([6] pmc.ncbi.nlm.nih.gov).

Pharma-specific "things to watch out for" therefore include:

- **Always verify MOA and mechanism statements** against textbooks or databases (e.g. DrugBank, PubChem). LLM outputs can be a first draft, but need expert correction.
- **Never trust LLM-generated clinical data or stats**. Simulated study results should be viewed as fiction unless backed by real sources.
- **Double-check all references**. If ChatGPT cites a guideline or study, manually look it up. In fact, it often fabricates them entirely ([6] pmc.ncbi.nlm.nih.gov) ([24] news.ashp.org).
- **Use specialized tools for interactions and dosing**. There are established DDI checkers and conversion calculators; rely on them instead of free-form text answers.
- **Be mindful of context and limitations**. For example, ChatGPT 3.5's knowledge stops in 2021 ([8] www.cnbc.com). Always confirm if new drugs (2022+) or updated regulations are involved.
- **Document AI contributions**. In regulated settings, maintain logs of AI prompts/outputs and vetting steps, to ensure traceability.
- **Maintain human oversight**. As many authors note, LLMs can only "augment" human experts, not replace them ([19] pmc.ncbi.nlm.nih.gov) ([44] pmc.ncbi.nlm.nih.gov). The final say in any drug-related decision must be a qualified professional.

Figure 1 illustrates the relative risk: in pharma, an AI hallucination can mean severe legal and health consequences, in stark contrast to benign errors in casual queries. (Note: This is a conceptual diagram.)

Pharma Context	LLM Hallucination Type	Potential Impact
Drug mechanism/explanations	Fabricated or incomplete mechanism	Misleads research or clinical understanding; could propagate false models ([14] pmc.ncbi.nlm.nih.gov)
Clinical trial data/results	Invented study data/results	False conclusions in research; fraudulent submissions (revistapesquisa.fapesp.br) ([1] jamanetwork.com)
Drug–drug or drug–herb interactions	Omitted/incorrect interaction(s)	Patient harm from unrecognized contraindications ([32] news.ashp.org) ([29] pmc.ncbi.nlm.nih.gov)
Dosage/regimen recommendations	Unit errors or miscalculated dose	Severe under/overdosing (e.g. 1000× error ([5] www.cnbc.com)); withdrawal or toxicity
Regulatory guidance & warnings	Invented or misquoted guidelines	Non-compliance, patient safety risk (false reassurance/alarm) ([2] www.medscape.com)
Citations and references	Fabricated/inaccurate sources	Undermines credibility; disseminates false "evidence" ([6] pmc.ncbi.nlm.nih.gov) ([24] news.ashp.org)

Table 2. Examples of LLM hallucination types in pharma contexts and their possible consequences. Generic hallucinations become far more serious when medicines are involved.

Future Directions and Recommendations

Generative AI in pharma is an active area of development, with ongoing efforts to mitigate these risks. Several approaches are emerging:

- **Retrieval-Augmented Generation (RAG):** By combining LLMs with real-time, authoritative databases, one can force the model to cite evidence. For instance, linking to DrugBank or PubMed during the query can reduce hallucinations. However, implementing RAG in regulated settings requires secure, validated data pipelines.
- **Domain-specific LLMs:** Large models pre-trained on general web text can be fine-tuned on medical/pharma corpora (e.g. Clinical BERT, BioGPT). Specialized models have smaller vocabularies and better factual grounding in biomedical knowledge. Early work (Wu et al.) suggests pharma-oriented LLMs perform better on specialized queries ([18] pmc.ncbi.nlm.nih.gov). Future AI tools may rely on such models.
- **Human-in-the-Loop Workflows:** Instead of a "one-shot" ChatGPT answer, systems should involve iterative checking. For example, clinicians might ask ChatGPT for an outline, then use software or staff to verify each claim. The use of chain-of-thought prompting (forcing the LLM to reason step-by-step) or requiring it to provide source links (with multi-turn follow-ups) are band-aids, but experts caution these do *not* eliminate hallucinations ([45] pmc.ncbi.nlm.nih.gov).
- **AI Fact-Checking Tools:** Research is underway on automated detection of AI-generated falsehoods. Techniques compare a model's answer to known facts or check statistical anomalies (as done by Giannaccare et al. to spot their fake list (revistapesquisa.fapesp.br)). In pharma, specialized fact-checkers could flag inconsistency in a drug description or data.
- **Regulatory Guidance and Standards:** The FDA's recent draft guidance for AI in drug development ([7] www.fda.gov) formally acknowledges the risk of "unvalidated model outputs." Eventually, AI tools used in regulated submissions will likely require validation akin to a medical device. Clinicians and industry should monitor these guidances: compliance may soon demand documenting AI use and proving accuracy for each claimed application.
- **Role Clarification:** It is crucial to define what AI *should* and *should not* do in pharma. The Inovia blog puts it well: use AI for tactical tasks (organizing data, hypothesis generation) but not in place of critical judgment (blog.inovia.bio). For example, AI might help **search** the literature quickly, but a pharmacist should verify all findings from trusted databases. Training for professionals should cover AI literacy: knowing LLM strengths (writing assistance, quick summaries) and pitfalls (fabrications, outdated info).
- **Transparency and Disclosure:** When AI tools are used in research or documentation, journals and companies should disclose it. Experiences from 2023–2025 suggest this is still evolving. As a community, we may need guidelines like "no AI-generated content without human-authored editing" (some journals already ban unannotated AI text). Regulatory agencies might in future require clear labeling if an AI was used in generating, say, patient educational materials.
- **Continuous Monitoring:** The landscape is changing fast. New models (ChatGPT-4.5, Gemini, etc.) claim reduced hallucinations, and APIs can now browse current data. We should scan emerging literature (e.g. studies comparing new LLMs on bio tasks) and update risk assessments accordingly.

In summary, LLMs have tremendous potential to accelerate aspects of drug discovery, literature review, and even patient education. But that potential comes with new classes of risk. The pharmaceutical context magnifies the stakes of hallucinations. As one expert group notes, LLMs in medicine can "augment human capabilities" only when properly constrained ([19] pmc.ncbi.nlm.nih.gov). We conclude with a cautious note: focus on leveraging LLMs with **precision and verification**, and never as an unquestioned oracle.

Conclusion

In this comprehensive review, we have shown that LLM hallucinations in biotech/pharma are not just "cute AI quirks," but can be severe missteps implicating patient safety and scientific integrity. We documented multiple domains of failure:

- **Mechanism of action**: ChatGPT can omit or misidentify key targets, undermining drug understanding ([14] pmc.ncbi.nlm.nih.gov).
- **Clinical trial data**: GPT-4 readily fabricates statistical evidence for trials (revistapesquisa.fapesp.br), a form of fraud that could devastate research trust.

- **Drug interactions**: The model may ignore serious interactions (e.g. Paxlovid/verapamil ([32] news.ashp.org)) or not fully explain them ([46] pmc.ncbi.nlm.nih.gov) ([34] pmc.ncbi.nlm.nih.gov).
- **Dosage advice**: LLMs can make unit conversion errors (1000× off ([5] www.cnbc.com)) with obvious harm potential.
- **Regulatory info**: ChatGPT confidently invents drug label warnings and guidelines ([2] www.medscape.com).
- **Citations**: An overwhelming majority of references it cites are fictitious ([6] pmc.ncbi.nlm.nih.gov) ([24] news.ashp.org).

Pharmaceutical applications demand *full accuracy*, yet the best LLMs today fall short of that mark. A clinical mistake is measured in lives, regulatory violations in fines and legal action, and scientific fraud in wasted effort. Each of the cases above illustrates how easily ChatGPT's "impressive language" can mislead in these high-stakes arenas.

Thus, we strongly advise: whenever generative AI is used in drug research or healthcare, **critical human review** is mandatory. Treat LLM outputs as provisional drafts, not facts. Use specialized databases (FDA labels, PubMed, official dosing calculators) for any decision affecting patient care or research conclusions. Regulatory bodies are catching up (FDA's AI guidance ([7] www.fda.gov), ASHP's alert to practitioners ([43] news.ashp.org)), but practitioners themselves must remain vigilant.

Future work should continue to evaluate LLM performance on domain-specific benchmarks, improve alignment with biomedical knowledge, and develop methods for real-time hallucination detection. In the meantime, the examples collected here serve as a warning. **We must identify "pharma-specific things to watch out for":** each category of hallucination can be life-threatening or career-ending in pharma, far beyond the inconvenience of a wrong trivia fact.

Our analysis underscores a central principle: **validity over novelty**. An answer from an LLM is valuable only if it can be traced and verified. As drug development becomes more data-driven, AI will undoubtedly play a role. But until LLMs can be rigorously verified against true pharmacological knowledge, they should be used with extreme caution in biotech and healthcare.

References

1. Long Island University ASHP Abstract – Grossman (*study at ASHP 2023 meeting*). **[ASHP Press Release (Dec 5, 2023)]** *Study Finds ChatGPT Provides Inaccurate Responses to Drug Questions*. Retrieved via ASHP News ([24] news.ashp.org) ([32] news.ashp.org).
2. Hsu et al., *JMIR Med Educ*, 2023 – *Examining Real-World Medication Consultations and Drug-Herb Interactions (ChatGPT performance)* ([26] pmc.ncbi.nlm.nih.gov) ([28] pmc.ncbi.nlm.nih.gov).
3. Giannaccare et al., *JAMA Ophthalmology*, 2023 – *LLM ADA abuse to create fake dataset* ([20] jamanetwork.com) ([1] jamanetwork.com).
4. Taloni, Scorcia, Giannaccare; *JAMA Ophthalmology*; *published online Nov 9, 2023* (the letter cited by Nature).
5. Gronlund et al., *Cureus*, 2023 – *Study on ChatGPT and DDIs* ([47] pmc.ncbi.nlm.nih.gov).
6. Bhattacharyya et al., *Cureus*, 2023 – *High rates of fabricated references in ChatGPT content* ([6] pmc.ncbi.nlm.nih.gov).
7. Bernadette Cornelison et al., *Pharmacy (Basel)*, 2025 – *ChatGPT and OTC medication in pregnancy*.
8. Iannantuono et al., *PLOS One (immuno-oncology LLM study)*, 2024 – *Comparison of LLMs on IO questions* ([48] pmc.ncbi.nlm.nih.gov).
9. Brook et al., *Medscape News*, Oct 2023 – *AI chatbot 'Hallucinates' faulty medical intelligence* ([12] www.medscape.com) ([2] www.medscape.com).
10. Ribon et al., *Live Science*, Aug 2025 – *Case report: Bromide intoxication after ChatGPT diet advice* ([39] www.livescience.com) ([4] www.livescience.com).

11. ChatGPT Usage Policies – OpenAI Terms (2023).
12. FDA AI Draft Guidance, Jan 2025 – *Considerations for AI in Regulatory Decision-Making* ([7] www.fda.gov).
13. Inovia Blog, Jan 2025 – *The AI Paradox in Pharma: Hype vs Reality* (blog.inovia.bio).
14. NTI (Nature News), Nov 22, 2023 – *ChatGPT generates fake data set to support hypothesis* ([22] www.nature.com).

External Sources

[1] https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2811505#:~:The%2...
[2] https://www.medscape.com/viewarticle/997510#:~:For%2...
[3] https://www.medscape.com/viewarticle/997510#:~:antib...
[4] https://www.livescience.com/health/food-diet/man-sought-diet-advice-from-chatgpt-and-ended-up-with-bromide-intoxication#:~:The%2...
[5] https://www.cnbc.com/2023/12/05/free-chatgpt-may-incorrectly-answer-drug-questions-study-says.html#:~:But%2...
[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC10277170/#:~:Overa...
[7] https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions#:~:A%20k...
[8] https://www.cnbc.com/2023/12/05/free-chatgpt-may-incorrectly-answer-drug-questions-study-says.html#:~:Notab...
[9] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:Many%...
[10] https://www.medscape.com/viewarticle/997510#:~:The%2...
[11] https://www.medscape.com/viewarticle/997510#:~:Artif...
[12] https://www.medscape.com/viewarticle/997510#:~:ChatG...
[13] https://www.livescience.com/health/food-diet/man-sought-diet-advice-from-chatgpt-and-ended-up-with-bromide-intoxication#:~:Live%...
[14] https://pmc.ncbi.nlm.nih.gov/articles/PMC10705618/#:~:%280%...
[15] https://pmc.ncbi.nlm.nih.gov/articles/PMC10705618/#:~:betwe...
[16] https://pmc.ncbi.nlm.nih.gov/articles/PMC11802629/#:~:Accor...
[17] https://www.nature.com/articles/s41537-023-00379-4#:~:falsi...
[18] https://pmc.ncbi.nlm.nih.gov/articles/PMC11802629/#:~:pharm...
[19] https://pmc.ncbi.nlm.nih.gov/articles/PMC12235426/#:~:Meta,...
[20] https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2811505#:~:The%2...
[21] https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2811505#:~:BSCVA...
[22] https://www.nature.com/articles/d41586-023-03635-w#:~:Resea...
[23] https://www.medscape.com/viewarticle/997510#:~:ChatG...
[24] https://news.ashp.org/News/ashp-news/2023/12/05/study-finds-chatgpt-provides-inaccurate-responses-to-drug-questions#:~:Only%...

[25] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02824-5/tables/6#:~:,Use%...

[26] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:Resul...

[27] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:,heal...

[28] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:hospi...

[29] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:inacc...

[30] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:Table...

[31] https://pmc.ncbi.nlm.nih.gov/articles/PMC10477918/#:~:Concl...

[32] https://news.ashp.org/News/ashp-news/2023/12/05/study-finds-chatgpt-provides-inaccurate-responses-to-drug-questions#:~:In%20...

[33] https://news.ashp.org/News/ashp-news/2023/12/05/study-finds-chatgpt-provides-inaccurate-responses-to-drug-questions#:~:%E2%8...

[34] https://pmc.ncbi.nlm.nih.gov/articles/PMC10105894/#:~:Among...

[35] https://pmc.ncbi.nlm.nih.gov/articles/PMC10105894/#:~:ChatG...

[36] https://www.cnbc.com/2023/12/05/free-chatgpt-may-incorrectly-answer-drug-questions-study-says.html#:~:Pharm...

[37] https://www.cnbc.com/2023/12/05/free-chatgpt-may-incorrectly-answer-drug-questions-study-says.html#:~:conve...

[38] https://www.medicaleconomics.com/view/ai-chatbots-lack-skepticism-repeat-and-expand-on-user-fed-medical-misinformation#:~:AI%20...

[39] https://www.livescience.com/health/food-diet/man-sought-diet-advice-from-chatgpt-and-ended-up-with-bromide-intoxication#:~:A%20m...

[40] https://www.livescience.com/health/food-diet/man-sought-diet-advice-from-chatgpt-and-ended-up-with-bromide-intoxication#:~:Recov...

[41] https://www.livescience.com/health/food-diet/man-sought-diet-advice-from-chatgpt-and-ended-up-with-bromide-intoxication#:~:of%20...

[42] https://www.livescience.com/health/food-diet/man-sought-diet-advice-from-chatgpt-and-ended-up-with-bromide-intoxication#:~:unkno...

[43] https://news.ashp.org/News/ashp-news/2023/12/05/study-finds-chatgpt-provides-inaccurate-responses-to-drug-questions#:~:%E2%8...

[44] https://pmc.ncbi.nlm.nih.gov/articles/PMC12389367/#:~:indis...

[45] https://pmc.ncbi.nlm.nih.gov/articles/PMC12235426/#:~:Large...

[46] https://pmc.ncbi.nlm.nih.gov/articles/PMC10105894/#:~:Among...

[47] https://pmc.ncbi.nlm.nih.gov/articles/PMC10105894/#:~:Resul...

[48] https://pmc.ncbi.nlm.nih.gov/articles/PMC10705618/#:~:the%2...

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.