JUNCTURE RESEARCH · REPORT

The Off-Label Drift Report

How AI answer engines rewrite approved pharma claims, and the regulatory exposure it creates.

JUNE 2026

JUNCTURE.HEALTH

Executive summary

A clinician with a question about a therapy no longer opens the label. They open an answer engine and type the question in plain language, and a fluent paragraph comes back. That paragraph was never written by anyone at the brand. It was reconstructed on demand by a system that retrieved a handful of sources and summarized them, and the summary dropped whatever clauses read as spare detail. The clauses that read as spare detail are the ones that do regulatory work: the qualifier that scopes the indication, the contraindication that carries fair balance. When they are dropped, an approved claim reads off-label, and the off-label reading is spoken to a prescriber with the full confidence of the writing style, with no human in the loop to catch it.

This report is for the people accountable for that message: regulatory, medical, and compliance leaders. It makes four arguments. First, that off-label drift is structural, a predictable output of how retrieve-then-summarize systems work, not a bug a vendor will patch. Second, that the clauses these systems drop first are precisely the regulated ones. Third, that the resulting answers map cleanly onto existing prohibitions in 21 CFR 202.1, in an enforcement climate that turned sharply stricter in 2025. Fourth, that the exposure is measurable, and that measuring it continuously is the only honest alternative to discovering it in a forwarded screenshot. The report closes with a three-move action plan and a description of how Juncture's Answer Monitor instruments the measurement.

The shift: clinicians have moved to the answer box

The premise behind every downstream argument is that the audience for your label has changed where it goes for an answer. The evidence is no longer thin.

In the Doximity 2026 State of AI in Medicine report, 63 percent of surveyed physicians reported currently using AI in their clinical practice, up from 47 percent in the survey a year earlier, and literature search was the single most common use case, rising to 35 percent from 22 percent. Literature search is the polite name for the behavior that matters here: a clinician asking a machine to summarize what is known about a drug. A 2024 Wolters Kluwer survey found that more than two-thirds of U.S. physicians (68 percent) had come to view generative AI as beneficial to

healthcare, a sentiment shift that runs ahead of any policy governing the use. The pattern is not confined to the United States. An updated 2025 survey of UK general practitioners found that one in four (25 percent) reported purposefully using generative AI tools in clinical practice, up from 20 percent the year before.

Two things follow. The adoption curve is steep and still rising, so the surface keeps growing. And the dominant use is exactly the one that exposes your claim: not drafting a note, but asking the machine what a drug does. The clinician is no longer reading your sentence. They are reading the machine's reconstruction of it.

The mechanism: retrieval and summarization both betray the label

To see why a faithful source can still produce an off-label answer, follow how the answer is actually built. There are two stages, and each is lossy in a way that works against a regulated claim.

The first stage is retrieval. When a clinician asks about a drug, the engine embeds the question and pulls the passages most similar to it from an index. Similarity is not completeness. A safety paragraph and an efficacy paragraph from the same label score very differently against a question like "what is this used for," and the efficacy passage wins. The contraindication never enters the model's context, so the model cannot include what it never received. A 2025 survey of retrieval-augmented generation architectures catalogs exactly these retrieval-side failures: noisy passages, partial coverage, and the model extrapolating beyond what the retrieved evidence supports (arXiv, 2025).

The second stage is summarization, and this is where the damage is done. Even when the right passages are retrieved, the model must compress them into a short answer, and abstractive summarization is lossy compression optimized for fluency. Fluency favors the main clause over the subordinate one. Research on multi-document summarization finds that state-of-the-art models hallucinate and drop content precisely when sources are diverse or partially overlapping (ACL Findings, 2025), and faithfulness degrades further as the material grows longer (arXiv, 2025). A qualifier is, structurally, the most droppable token in a sentence: it is subordinate,

it is conditional, and removing it makes the prose read cleaner. The model is optimizing for the thing that erases your protection.

A common assumption is that a cited answer is a safe answer. It is not. Work formalizing citation faithfulness shows that a model often produces a claim from its own parametric memory and then attaches a citation that maps to a document only superficially, a phenomenon the authors call post-rationalization (Wallat et al., 2024). Correctness of the citation and faithfulness of the claim to that citation are different properties. A clinician who clicks the link and sees a real label may trust a sentence the label never contained.

Watching the mechanism drift a Varigel claim

Make it concrete. Varigel is a fictional therapy approved for one narrow indication, with a contraindication in patients taking a common comorbidity medication. As cleared by MLR, the approved claim reads in full:

> **Varigel is indicated for moderate to severe Condition X in adults who have not responded to first-line therapy. Varigel is contraindicated in patients receiving Medication Y.**

Two clauses do regulatory work. "Who have not responded to first-line therapy" scopes the indication. "Contraindicated in patients receiving Medication Y" is fair-balance safety information. Now run the pipeline.

Retrieval. The clinician asks "what is Varigel used for." The retriever scores passages by similarity to that question. The indication paragraph scores high. The contraindication paragraph, which never says "used for" and reads as safety prose, scores lower and falls below the cutoff. It is not retrieved. Before the model writes a token, the safety clause is gone from the context.

Summarization. The model now holds the indication paragraph plus a competitor-comparison blog and a years-old conference abstract that also surfaced. It

compresses. "Moderate to severe Condition X in adults who have not responded to first-line therapy" is long and conditional, so the model renders it as "Varigel treats Condition X in adults." The qualifier is dropped, not maliciously, but because the shorter clause is more fluent. The abstract mentioned an exploratory second use, so the model, reaching to complete the answer, adds "and is also used for Condition Z."

The output:

> **Varigel is used to treat Condition X in adults, and is also used for Condition Z.**

Compare it to the approved claim. The first-line-failure qualifier is gone, so the indication has silently broadened. The contraindication is gone, so fair balance is gone. An exploratory mention has been promoted to "used for," so the answer reads off-label. Three regulated transformations, zero invented facts. Every word traces to a real source. The drift lives entirely in what was selected and what was compressed away.

This is not a thought experiment. Studies of ChatGPT answering real drug-information questions repeatedly find responses that are incomplete or only partially correct and that require careful human review before use (**iForumRx, 2024**; **JACCP, 2025**). Incomplete is the operative word. The failure is usually omission, the dropped clause, not fabrication.

Why this is structural, not a one-off

The strongest evidence that drift is built in comes from the domain that has worked hardest to engineer it out. Stanford researchers tested purpose-built legal research tools that use retrieval specifically to ground answers in authoritative sources. They still produced incorrect or misgrounded answers on roughly one in six queries, with rates of 17 percent and higher, and the errors included citing a real source for a claim that source did not support (**Stanford HAI, 2024**). Retrieval lowered the error rate. It did not remove it. If a tool engineered for a regulated, citation-obsessed profession

misgrounds one answer in six, a general engine paraphrasing your label is not going to do better.

Three properties make the drift recur rather than resolve. It is non-deterministic: the same question, asked twice or asked on two engines, retrieves a slightly different passage set and samples a slightly different summary, so a single screenshot certifies nothing. The droppable token is the regulated one: qualifiers, indications, and contraindications are exactly the subordinate clauses summarization sheds first. And silence is filled, not respected: when retrieval undercovers, the model reaches for ambient third-party content to complete the paragraph rather than returning a shorter, safer answer. A brand that never published a clean, machine-legible version of its claim is not met with silence. It is met with a confident reconstruction assembled from whatever was nearby.

The regulatory exposure

Reframe the Varigel output as a promotional artifact and the problem sharpens. FDA's prescription-drug advertising rule requires a fair balance of benefit and risk information and prohibits promotion of a drug for an unapproved use (21 CFR 202.1). The drifted answer dropped the contraindication, so risk information is no longer presented alongside benefit. It promoted an exploratory use to an indication, so it reads as off-label promotion. The rule ties promotion to uses for which the drug is generally recognized as safe and effective, supported by adequate and well-controlled investigations (eCFR). The drifted "also used for Condition Z" clears none of that.

The 2025 enforcement climate makes this more than a theoretical mapping. On September 9, 2025, FDA's Office of Prescription Drug Promotion sent 46 untitled letters in a single day, more than double the number it had sent in the previous five years combined, and closed the year with 74 letters to drug and biologic manufacturers, 42 of them aimed at direct-to-consumer television ads. The dominant theme across the wave was omission and minimization of risk information (Covington, 2025), and the letters repeatedly cited misleading efficacy conveyed through visual presentation, including exaggerated quality-of-life improvement and implied complete resolution of symptoms (King & Spalding, 2025). A drifted answer

that drops the contraindication is the same failure, omission of risk information, produced by a machine instead of an agency.

The uncomfortable part is responsibility. In two of the 2025 letters the FDA took an aggressive view of who is responsible for a promotional communication. If your brand fed or influenced the source the engine leaned on, the line from your content to the off-label sentence is short. And the deeper exposure is ownership inside your own house. MLR governs what you publish. Medical affairs governs the evidence. Neither has a standing mandate over a sentence a third-party model generates when no one from your company is in the room. The drift is a regulated communication that your governance process never saw, because your governance process was built to review documents, and this is not a document. It is a reconstruction, produced on demand, different every time.

A measurement framework

You cannot govern what you cannot see, and you cannot see a per-user reconstruction by reading the label. The instrument has to point at the output. Two measures, run together and run continuously, turn the invisible surface into a managed one.

The first is Share of Answer. For the real questions your audience asks, in their words, how often is your brand present in the answer at all, how prominently, and against which competitors. Share of Answer is the visibility baseline. A brand absent from the answer is not safe, it is undefended, because the model will still answer the question using whatever ambient content it found.

The second is off-label drift detection. For every answer where the brand is present, the system compares the machine's sentence against the approved claim and flags three failure modes: a dropped or broadened qualifier, a missing or minimized contraindication or other safety clause, and an indication that reaches beyond the approved label. The flag is only useful if it cites the clause. "This answer reads off-label" is an anxiety. "This answer dropped the first-line-failure qualifier and omitted the Medication Y contraindication, measured against label section X" is a work item a regulatory reviewer can act on.

Three properties make the framework sound rather than decorative. It runs across the five engines clinicians actually use, ChatGPT, Gemini, Perplexity, Google AI Overviews, and Claude, because drift differs by engine and a result on one says nothing about the others. (OpenEvidence and other clinical answer engines are a planned area to extend into, not a currently monitored surface.) It is continuous, because the output is non-deterministic and moves when a model updates, an index refreshes, or a competitor publishes, so a baseline taken once is a screenshot of a river. And it is anchored to a known-good claim, so a machine answer is read not as a free-floating paragraph but as a measured deviation from the exact sentence you cleared.

A three-move action plan

You do not need a moonshot. You need to treat the machine's answer as a measurable surface and put three things in place.

1. Take the baseline before you do anything else. Pick the twenty questions your HCPs and patients actually ask, in their words, not your campaign headlines. Run them across ChatGPT, Gemini, Perplexity, Google AI Overviews, and Claude. Record three things per answer: are you mentioned, does the mention match the label, and does anything read off-label. That is your exposure, in numbers, for the first time.

2. Trace every off-label reading back to a source. A drift you cannot source is a drift you cannot fix. For each off-label or omitted-safety answer, find what the model leaned on: the stale abstract, the third-party summary, the gap your own content left open. Sourcing the drift turns a vague anxiety into a closable work item, and it tells you whether the fix is to publish a cleaner approved source or to correct something already in the wild. Give the pipeline a clean source to prefer: your approved indication, your contraindication, and your safety language, published as plain, well-structured, machine-legible content, are more likely to be retrieved whole and compressed faithfully than a clause buried in a PDF.

3. Watch it continuously and route it back to MLR. When a model updates, when a competitor publishes, when a new abstract surfaces, the answer moves. The brands that get ahead of this will detect a new off-label reading the week it appears, trace it

to its source, and route it to the people who own the underlying content, so the correction lands in the system of record and not in a panicked email thread.

How Juncture instruments it

The reason most brands cannot run those three moves is that the inside and the outside have never been connected. The team that approves the message has no view of what the machine says. Whoever might watch the machine has no authority over the approved message. So drift is found late, sourced never, and corrected by accident.

Juncture is built for that seam. Inside, it pre-checks the approved message before MLR, comparing each asset against the label, surfacing how much of it reuses already-approved content, and backing the reviewer with a 21 CFR Part 11-supporting audit trail and an e-signature sign-off. Outside, its Answer Monitor measures Share of Answer and detects off-label drift across the five engines your audience uses, continuously, and traces each drift back to the label clause it violated.

The value is the join. Because Juncture already holds the approved sentence you cleared on the inside, it reads a machine answer not as a free-floating paragraph but as a deviation from a known-good source: which off-label reading, which dropped qualifier, which missing contraindication, measured against the exact clause that should have been there. A dropped contraindication or an off-label "used for" surfaces as a tracked exception with a citation, not a surprise in a forwarded screenshot. The wording on the inside matters and is deliberate: the tool is Part 11-supporting, supplying the technical controls, while the customer validates the system under their own SOPs and owns the compliance posture.

The takeaway

Off-label drift is not the model lying. It is retrieval selecting an incomplete slice of your message and summarization compressing away the qualifiers and contraindications that made it on-label. Both stages work as designed, which is why patching one bad answer does nothing: the next query runs the same lossy pipeline

and produces the next drift. The audience has moved to the answer box, the enforcement climate has tightened, and the regulated clauses are precisely the ones the machine drops first.

The off-label answer is already out there. The only question is whether you see it as a measured deviation from a sentence you control, or hear about it in a screenshot after a prescriber has already read it. Bring one brand and the twenty questions your audience actually asks. We will show you what the machine says about it today, flag the off-label drift already in the wild, and trace each one back to the approved clause it broke.

Sources

1. Doximity, "2026 State of AI in Medicine Report," 2026. doximity.com
2. Wolters Kluwer, "Over two-thirds of U.S. physicians have changed their mind, now viewing GenAI as beneficial in healthcare," 2024. wolterskluwer.com
3. "General practitioners' adoption of generative artificial intelligence in clinical practice in the UK: an updated online survey," PMC, 2025. pmc.ncbi.nlm.nih.gov
4. Gupta et al., "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," arXiv, 2025. arxiv.org
5. "How LLMs Hallucinate in Multi-Document Summarization," ACL Findings (NAACL), 2025. aclanthology.org
6. "Hallucinate at the Last in Long Response Generation: A Case Study on Long Document Summarization," arXiv, 2025. arxiv.org
7. Wallat et al., "Correctness is not Faithfulness in Retrieval Augmented Generation Attributions," arXiv, 2024. arxiv.org
8. Magnus et al., "Helpful or Harmful? Using ChatGPT to Answer Drug Information Questions," iForumRx, 2024. iforumrx.org
9. Khatri et al., "Accuracy and reproducibility of ChatGPT responses to real-world drug information questions," JACCP, 2025. accpjournals.onlinelibrary.wiley.com

0. Magesh et al., "AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries," Stanford HAI, 2024. hai.stanford.edu
1. Legal Information Institute, "21 CFR 202.1 Prescription-drug advertisements." law.cornell.edu
2. Electronic Code of Federal Regulations, "21 CFR 202.1 Prescription-drug advertisements." ecfr.gov
3. Sheppard Mullin, "FDA's Wave of Untitled Letters Signals Stricter Scrutiny for DTC Pharma Ads," October 2025. sheppard.com
4. Covington & Burling, "FDA Advertising and Promotion Enforcement Activities: Update," October 2025. cov.com
5. King & Spalding, "2025 Year in Review: FDA Drug and Device Advertising and Promotion Enforcement," 2025. kslaw.com
6. Latham & Watkins, "FDA Begins Crackdown on Direct-to-Consumer Pharmaceutical Advertising," September 2025. lw.com