JUNCTURE RESEARCH · WHITEPAPER

The Six Core KPIs for Pharma Share of Answer

A label-grounded measurement framework for how ChatGPT, Gemini, Perplexity, Google AI Overviews, and Claude answer the questions HCPs actually ask.

JUNE 2026

JUNCTURE.HEALTH

Why share of voice fails in the answer era

For twenty years, digital pharma measurement rested on a proxy: if you could count impressions, placements, and rank positions, you could infer mind share. Share of voice was that proxy. It worked because the search results page was a list, and a list is a competition for clicks you could observe and buy into.

Answer engines dissolve the list. The user asks a question in natural language and receives a synthesized answer, often with no obligation to click anything. The scale of this shift in health is not speculative. In an analysis of 130,070 U.S. health-related search queries collected in July 2025, AI Overviews appeared at a 51.6% rate, the highest of any industry studied and roughly double the average across industries (WebFX). A separate longitudinal analysis found AI Overview presence on symptoms-and-conditions queries rising to 93% and on treatment-and-procedure queries to 100% by late 2025 (BrightEdge).

The behavioral consequence is measured. The Pew Research Center, analyzing the browsing data of 900 U.S. adults during March 2025, found that users who encountered an AI summary clicked a traditional search result only about 8% of the time, versus 15% when no summary appeared, and that they clicked a link inside the summary itself in just 1% of visits (Pew Research Center). The answer is now the endpoint of the journey.

That is fatal for share of voice. The metric measures the competition for a click that, in the majority of health queries, no longer happens. It tells you nothing about the content of the answer the model gave: whether it named your brand, whether the dosing it stated matched your label, whether it surfaced an off-label use, or whether it cited a payer policy page instead of your prescribing information. A 2024 online survey of 1,006 UK general practitioners found that 20% already reported using generative AI tools in clinical practice, with the most common uses being generating documentation after appointments and suggesting a differential diagnosis (Blease et al., BMJ Health & Care Informatics). The audience is inside the engine. Pharma needs a metric that measures what the engine says, not what it displays.

The denominator problem: what are we even measuring?

Every share metric is a fraction, and the integrity of the fraction lives in the denominator. Share of voice had a clean denominator: a fixed set of ad slots or a ranked results page. Share of Answer has no such gift. There is no page to count. The output is generated fresh, varies by phrasing, varies by engine, and varies over time. If you do not define the universe of questions you are measuring against, your numerator is meaningless.

The wrong denominator is the vanity prompt. It is tempting to ask each engine "what is the best treatment for condition X" and score the result. But that is not what HCPs ask, and it produces a number that flatters or alarms with no decision value. The right denominator is a governed, versioned question set drawn from how clinicians and informed patients actually phrase their needs.

A defensible HCP question set is built from real signal, not invention. Sources that hold up to scrutiny include: de-identified medical information call-center and inquiry logs; field medical and MSL interaction summaries; the long-tail query data your own properties already capture; published clinical FAQs; and the question structures that answer engines themselves expose. These are then organized along the dimensions that matter clinically: dosing and administration, contraindications and warnings, drug interactions, comparative efficacy, mechanism of action, special populations, and access. Each question is tagged by intent and by therapeutic relevance, and the set is frozen as a versioned artifact so that a movement in the metric reflects a movement in the engines, not a change in what you asked.

Three rules keep the denominator honest. First, it must be representative, weighted toward the high-volume, high-stakes questions an HCP genuinely raises, not the questions that make your brand look good. Second, it must be stable and versioned, because a share trend computed against a shifting question set is noise. Third, it must be label-anchored, meaning every question maps to a region of your approved label so the answer can be graded against an authoritative ground truth rather than against a competitor's marketing. With the denominator fixed, the six KPIs become computable.

The first three KPIs: presence, breadth, and accuracy

Share of Answer (SoA) is the foundation. For a defined question set and a defined engine, Share of Answer is the proportion of answers in which your brand is mentioned, out of all answers where any in-class brand or your brand could reasonably appear. If Varigel and its competitors are eligible to be named across 200 dosing-and-efficacy questions, and Varigel is named in 96 of the engine's answers, its raw presence is 96. Share of Answer expresses that against the eligible field, so a brand that is named in 96 answers while the whole class is named in 240 brand-mentions holds a 40% Share of Answer. It is the direct successor to share of voice, but the unit counted is a mention inside a synthesized answer to a real clinical question, not an impression.

Ecosystem Share of Answer (ESoA) corrects a blind spot in raw SoA. Engines do not answer in a vacuum; they answer relative to a class. Ecosystem Share of Answer measures your presence against the full competitive set as the engines construct it, including comparators you might not consider direct rivals and including the generic or class-level framing models often default to. A brand can hold a healthy raw SoA while the engine consistently frames the class around a competitor's positioning. ESoA exposes that. It answers the question, "when the model talks about this therapeutic area, how much of the conversation is us, and who are we actually being compared against." This matters because answer engines synthesize a category narrative, and being absent from that narrative is a strategic loss that raw presence counting hides.

Precision of Answer (PoA) is the accuracy KPI, and it is where label grounding becomes non-negotiable. Precision of Answer measures, among the answers that mention your brand, the proportion of factual claims that are correct and consistent with your approved label. This is not sentiment and not tone. It is a claim-by-claim comparison of what the model asserted against the label-grounded ground truth: the dose it stated, the indication it described, the population it named, the safety language it used. The need is concrete. A study evaluating large language models on critical-care pharmacotherapy assessments found a top model accuracy of 71.6% with a baseline prompt, rising to 77.4% with few-shot chain-of-thought prompting, and the models performed markedly worse on knowledge-application questions than

on simple recall, with the strongest model scoring 67% on application versus 87% on recall (Yang et al., PMC). Roughly a third of application-level medication answers being wrong is not a rounding error for a regulated product. Precision of Answer turns that risk into a tracked, auditable number.

The second three KPIs: risk, language uptake, and trust

Risk of Answer (RoA) is the safety and compliance KPI, and it is the inverse lens on the same evidence Precision of Answer uses. Where Precision asks "how much was right," Risk asks "how dangerous is what was wrong, and where." Risk of Answer is a weighted measure of the answers that contain an off-label claim, an omitted or understated safety statement, a contraindication error, an unsupported comparative superiority claim, or a fabricated reference. The weighting matters: a model that omits a boxed-warning-level safety statement is a different category of problem than one that rounds a half-life. Large language models are documented to generate plausible but incorrect or unverified information, a failure mode the research literature flags as carrying serious consequences in healthcare applications (Pal et al., Med-HALT, arXiv). Risk of Answer makes the off-label and the fabricated answer a counted, escalatable event rather than a thing someone happens to notice.

Claim Uptake measures whether your approved language is the language the models echo. Every regulated brand maintains a library of approved claims, the specific, label-supported statements MLR has cleared. Claim Uptake is the proportion of your priority approved claims that actually appear, in substance, in the engines' answers. It is the bridge between the content you produce and the content the models reproduce. High Claim Uptake means your MLR-approved framing is winning the synthesis. Low Claim Uptake means the models are constructing your brand's story from someone else's words, which is simultaneously a marketing loss and a compliance exposure. The foundational generative-engine-optimization research is instructive here: structuring content with clear, source-citable, quotation-friendly statements measurably increased visibility in generative answers, by up to 40% in the original study (Aggarwal et al., GEO, KDD 2024). Claim Uptake measures the result of that work on the claims you care about.

Top References answers "whose content does the engine trust." For each question and engine, you capture the sources the model cites or draws on, then aggregate to see which domains repeatedly inform answers about your brand and your class. This is the most actionable KPI for content strategy, because it tells you where the models are sourcing their understanding. If a third-party drug database, a payer policy page, or a competitor's site is consistently in the Top References while your prescribing information is absent, you know precisely where the synthesis is being shaped, and where corrective, label-grounded content needs to live to be picked up.

A worked example: Varigel across the five engines

Consider Varigel, a fictional specialty product. The brand team builds a versioned question set of 200 HCP questions across dosing, contraindications, interactions, comparative efficacy, and special populations, each mapped to a region of the Varigel label. They run the set against the five canonical engines, ChatGPT, Gemini, Perplexity, Google AI Overviews, and Claude, with a cadence calibrated to each engine's volatility: weekly sampling for the fastest-moving conversational engines, monthly for AI Overviews, and full re-baselining quarterly. Each question is sampled multiple times per engine to capture run-to-run variance, and the answers are graded by Answer Monitor against the label.

The first month's reading is sobering and specific. **Share of Answer** comes in at 40%: Varigel is named in 96 of the eligible brand-mentions across the class. But **Ecosystem SoA** reveals that two-thirds of the class narrative is built around a single competitor, and that the engines repeatedly frame Varigel as a "second-line alternative," a framing that appears nowhere in its label. **Precision of Answer** is 82%: of the factual claims in Varigel answers, 18% are wrong, concentrated in dosing for renal-impaired patients, exactly the special-population region the label addresses in detail. **Risk of Answer** flags 7 answers across the engines that imply an off-label maintenance use, and two that omit a key interaction warning. **Claim Uptake** shows that only 5 of the brand's 14 priority approved claims are echoed by the models. **Top References** explains why: a third-party medication database and two clinician-forum threads dominate the citations, while the Varigel prescribing information appears in the reference set for fewer than one answer in ten.

The framework converts this into governed action. The off-label maintenance-use answers and the omitted interaction warning become a logged Risk of Answer escalation for medical and regulatory review, time-stamped and reproducible, the kind of record a 21 CFR Part 11-supporting system is built to hold. The renal-dosing precision gap and the low Claim Uptake become a content brief: publish label-grounded, quotation-ready renal-dosing content and reinforce the five missing approved claims where the Top References data says the engines are actually looking. The "second-line alternative" misframing becomes an Ecosystem SoA objective. Next quarter, the same frozen question set re-run against the same five engines shows whether the interventions moved the numbers. That closed loop, real questions in, graded answers out, governed action, re-measurement, is the difference between a dashboard and a measurement system.

Where this leaves you

Share of voice measured a competition that answer engines have ended. In a category where the majority of health queries now return an AI-synthesized answer and users rarely click through, the question that matters is no longer "did my placement appear" but "what did the model tell the physician, was it on-label, and whose content did it trust." The six Core KPIs answer that question with rigor: Share of Answer and Ecosystem SoA for visibility, Precision of Answer and Risk of Answer for accuracy and safety, Claim Uptake for whether your approved language wins the synthesis, and Top References for where the engines source their understanding. None of these are computable without two disciplines: a versioned, representative HCP question set as the denominator, and a label as the ground truth.

This is the discipline Juncture Answer Monitor operationalizes. It runs a governed question set against ChatGPT, Gemini, Perplexity, Google AI Overviews, and Claude on a per-engine cadence, headlines the six Core KPIs, grades every claim against the approved label, flags off-label and drift events, and shows the Top References shaping the answers. Because it shares a claims library and label spine with Pre-Check and Content Intelligence, the inside-out join is direct: the approved claim you cleared in Pre-Check is the same claim whose uptake Answer Monitor measures in the wild, and the content gap it surfaces is the brief Content Intelligence turns into

compliant, citable assets. That loop, from what you approved to what the models echo, is the measurement system the answer era requires.

Sources

1. Pew Research Center, "Google users are less likely to click on links when an AI summary appears in the results" (July 2025)
2. WebFX, "AI Overviews in Healthcare: What Our Study of 130K+ Health Queries Reveals" (2025)
3. BrightEdge, "Healthcare and AI Overviews: How Google Sharpened Its Approach Over Three Years" (2025)
4. Aggarwal, Murahari, Rajpurohit, Kalyan, Narasimhan, Deshpande, "GEO: Generative Engine Optimization," KDD 2024 (arXiv:2311.09735)
5. Blease, Locher, Gaab, Hägglund, Mandl, "Generative artificial intelligence in primary care: an online survey of UK general practitioners," BMJ Health & Care Informatics (2024) (PMC11429366)
6. Yang, Hu, Most, Hawkins, Murray, Smith, Li, Sikora, "Evaluating accuracy and reproducibility of large language model performance on critical care assessments in pharmacy education" (PMC11754395)
7. Pal, Umapathi, Sankarasubbu, "Med-HALT: Medical Domain Hallucination Test for Large Language Models" (arXiv:2307.15343)