

AI Chatbot Crisis-Referral “Compliance” Claims and the FTC Marketing-Fraud Question: An Evidentiary Review

TL;DR

- **The compliance percentages are documented facts, but the companies frame them as internal, self-graded model-behavior benchmarks — not as user-safety-outcome or clinical-efficacy claims.** OpenAI’s “91% compliant with our desired behaviors” (Oct 27, 2025) and Anthropic’s “respond appropriately 98.6%–99.3% of the time” (both from official company blog posts) describe adherence to the companies’ own policies on adversarially-selected synthetic test sets, self-scored by their own automated evaluators, and (in OpenAI’s case) explicitly labeled “not representative of average production traffic.” Named critics — ex-OpenAI safety lead Steven Adler and the *Raine v. OpenAI* complaint — have publicly characterized this self-grading and single-turn testing as potentially misleading.
- **There is real regulatory precedent for treating a percentage-accuracy AI claim as deceptive: the FTC’s April 2025 Workado action attacked a “98 percent” AI-accuracy claim as false/unsubstantiated under Section 5.** But no US or foreign regulator has yet taken any enforcement action specifically about AI *crisis-referral* or *suicide-prevention* efficacy/compliance claims. The FTC has an open September 2025 6(b) study of companion chatbots (covering all five companies), and Texas AG Paxton is investigating Meta/Character.AI for deceptively marketing bots as mental-health tools — but those target therapy-impersonation, not the compliance percentages.
- **A “marketing fraud” theory is legally coherent but currently speculative as applied to these specific percentages.** The FTC’s three-part deception test and its heightened “competent and reliable scientific evidence” health-claims standard could in principle apply — but the companies’ careful hedging (“compliant with our desired behaviors,” not “safe for users”) is precisely the framing that makes a deception case harder to establish.

Key Findings

1. How the metrics are actually worded and framed

- **OpenAI** (blog, “Strengthening ChatGPT’s responses in sensitive conversations,” Oct 27, 2025, filed under “Safety”): “On a model evaluation consisting of more than 1,000 challenging self harm and suicide conversations, our new automated evaluations score the new GPT-5 model at 91% compliant with our desired behaviors, compared to 77% for the previous GPT-5 model.” Parallel figures: 92% compliant on the mental-health taxonomy (vs. 27% for the prior GPT-5) and 97% on emotional reliance (vs. 50%). The post repeatedly frames the numbers as policy-adherence on adversarial internal evals, stating results “come from evaluations that are designed to not ‘saturate’ near perfect performance, and error rates are not representative of average production traffic,” and that “future measurements may not be directly comparable to past ones.”
- **Anthropic** (blog, “Protecting the wellbeing of our users”): “On requests involving clear risk, our latest models—Claude Opus 4.5, Sonnet 4.5, and Haiku 4.5—respond appropriately 98.6%, 98.7%, and 99.3% of the time, respectively. Our previous-generation frontier model, Claude Opus 4.1, scored 97.2%.” These are single-turn synthetic evaluations (“an individual message related to suicide or self-harm, without any prior conversation or context”). Anthropic also touts very low benign-refusal rates (0%–0.075%) and its ThroughLine partnership

providing crisis resources across “170+ countries,” including the 988 Lifeline in the US/Canada, the Samaritans in the UK, and Life Link in Japan.

- **Crucial framing point:** Neither company frames these numbers as claims about user safety outcomes or clinical efficacy. Both frame them as internal benchmark/policy-adherence metrics scored by their own systems. This is the central evidentiary distinction the question asks about. Neither figure was located in an SEC/regulatory filing or a formal investor communication; they appear in “Safety”-tagged blog posts and system-card addenda. (Note: for the same OpenAI evaluation set, the launch version of GPT-5 scored only 27% on the mental-health taxonomy — illustrating how far a self-created, self-graded number can swing depending on the test.)

2. Named critics who have called the metrics potentially misleading

- **Steven Adler** (former head of product safety at OpenAI, four years, departed 2024; now independent researcher, “Clear-Eyed AI” Substack). In a New York Times opinion piece (Oct 28, 2025), he wrote he has “major questions ... about whether these mental health issues are actually fixed,” [Futurism](#)
↗
and that “People deserve more than just a company’s word that it has addressed safety issues” — concluding, “Prove it.” On his Substack he argued that “when we let the AI companies grade their own safety work, we shouldn’t be surprised when they pass with flying colors,” [Substack](#)
↗
and that without independent auditing “companies grade their own homework.” [Substack](#)
↗
He also flagged that OpenAI reported prevalence data “without comparison to rates from the past few months,” [Futurism](#)
↗
calling the disclosure “a great first step” that did not go far enough.
- **Raine v. OpenAI** (First Amended Complaint, San Francisco Superior Court, case CGC-25-628528): alleges OpenAI “evaluated the model’s safety almost entirely through isolated, one-off prompts. By doing so, OpenAI not only manufactured the illusion of perfect safety scores, but actively concealed the very dangers built into the product it designed and marketed to consumers.” This directly attacks the single-turn methodology behind the percentages, pairing it with OpenAI’s own admission that “our safeguards work more reliably in common, short exchanges” [TechCrunch](#)
↗
and degrade in long conversations. OpenAI’s Nov. 25, 2025 answer denies liability, states GPT-4o “passed” its safety testing “as shown by the safety scorecard OpenAI publicly releases,” [Arstechnica](#)
↗
and argues the teen misused the product and bypassed guardrails.
- **Academic “safetywashing” literature** (Ren et al., “Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?”, NeurIPS 2024): a meta-analysis finding many safety benchmarks are highly correlated with general capabilities, so “advancements in safety may often simply reflect improvements in underlying model capabilities rather than meaningful progress in safety-specific research goals,” [AI Models](#)
↗
enabling capability gains to “masquerade as safety advances.” [ResearchGate](#)
↗
A broader European Commission-affiliated review (“Can We Trust AI Benchmarks?”, 2024) and MLCommons both document that safety benchmarks can be “misleading,” saturate over time, and suffer “low ecological validity.”
- **Independent peer-reviewed benchmark** (McBain et al., *Psychiatric Services*, published by the American Psychiatric Association, Aug. 26, 2025; conducted by RAND, funded by the National Institute of Mental Health; doi:10.1176/appi.ps.20250086): tested OpenAI’s GPT-4o mini, Anthropic’s Claude 3.5 Sonnet, and Google’s

Gemini 1.5 Pro, querying each 100 times across 30 clinician-rated suicide-related questions (roughly 9,000 responses, data collected Oct. 2024). It found the chatbots aligned with expert judgment at the extremes but “did not meaningfully distinguish intermediate risk levels ... the odds of a direct response were not statistically different for low-risk, medium-risk, or high-risk queries” (adjusted odds of a direct response: Claude aOR 2.01 and Gemini aOR 0.09 relative to ChatGPT). Notably, “ChatGPT and Claude sometimes answered questions about which firearms, poisons, or ropes had the highest rate of completed suicide,” which lead author Ryan McBain (RAND senior policy researcher and Harvard Medical School assistant professor) said “should have been treated as red flags.” This independent result — near-parity with experts only at the extremes, with meaningful failures in the middle — sits uneasily beside the companies’ near-perfect self-reported figures, though the two use different methods and question sets and are not directly comparable.

3. Regulatory and legal precedent

- **FTC v. Workado (April 2025 complaint; final order Aug. 2025) — the closest precedent.** The FTC alleged Workado’s “98 percent” accuracy claim for its “AI Content Detector” was “false, misleading, or non-substantiated.” [Federal Trade Commission](#)
↗
Per the FTC’s administrative complaint, the underlying model “struggled to identify AI-generated content as AI-generated when evaluating nonacademic content, correctly detecting AI-generated text merely 53.2% of the time”; [CyberScoop](#)
↗
Bureau of Consumer Protection Director Chris Mufarrige said the product “did no better than a coin toss.” [Federal Trade Commission](#)
↗
The consent order bars effectiveness representations unless non-misleading and backed by “competent and reliable evidence” at the time made, with retention and consumer-notification requirements. [Federal Trade Commission](#)
↗
This establishes that a percentage-accuracy claim about an AI system falls squarely within Section 5 deception authority and the substantiation doctrine.
- **FTC 6(b) study (Sept. 11, 2025):** orders to Alphabet, Character Technologies, Instagram, Meta Platforms, OpenAI, Snap, and xAI, seeking information on safety testing, monetization, and how companies “employ disclosures, advertising, and other representations.” [Federal Trade Commission](#)
↗
Commissioner Mark Meador’s accompanying statement invoked the death of Adam Raine and stated that “those who make [these chatbots] available have a responsibility to comply with the consumer protection laws.” [Federal Trade Commission](#)
↗
Important limit: 6(b) is a study authority “that does not have a specific law enforcement purpose,” [Federal Trade Commission](#)
↗
though the FTC can use findings to open investigations. (Separately, the FTC has reportedly been probing OpenAI’s consumer-protection compliance since 2023.)
- **Texas AG Ken Paxton (Aug. 18, 2025):** opened an investigation into Meta AI Studio and Character.AI for “potentially engaging in deceptive trade practices and misleadingly marketing themselves as mental health tools,” [TechCrunch](#)
↗
issuing Civil Investigative Demands. Paxton said the platforms, “By posing as sources of emotional support ... can mislead vulnerable users, especially children, into believing they’re receiving legitimate mental health care. In reality, they’re often being fed recycled, generic responses engineered to align with harvested personal data

and disguised as therapeutic advice.” [Office of the Attorney General](#)

↗

- **Other state actions:** Kentucky AG Russell Coleman filed the first state lawsuit against an AI chatbot company (Character Technologies, Jan. 8, 2026); Pennsylvania sued Character.AI (May 5, 2026) over bots posing as licensed psychiatrists (“We will not allow companies to deploy AI tools that mislead people into believing they are receiving advice from a licensed medical professional,” [NPR](#)

↗

said Gov. Shapiro); and a bipartisan coalition of 44 state AGs sent a warning letter to leading AI companies (Aug. 25, 2025). These are design-defect, child-safety, and therapy-impersonation theories — not challenges to the compliance percentages.

- **FTC deception and substantiation framework (the legal elements):** Under the 1983 Policy Statement on Deception, the FTC will find deception where there is (1) “a representation, omission or practice that is likely to mislead the consumer,” [KTS Law](#)

↗

(2) examined “from the perspective of a consumer acting reasonably in the circumstances,” [KTS Law](#)

↗

that is (3) “material” — i.e., “likely to affect the consumer’s conduct or decision” about the product. Intent is not required; a good-faith advertiser can still violate Section 5. Under the 1984 Advertising Substantiation Statement, express establishment claims (“tests prove,” “studies show,” a specific percentage) require the advertiser to possess at least the advertised level of support; absent such a claim, consumers are assumed to expect a “reasonable basis.” Under the FTC’s 2022 Health Products Compliance Guidance, claims about “the efficacy or safety of health-related products” [Federal Trade Commission](#)

↗

require “competent and reliable scientific evidence,” which “as a general matter ... will need to be in the form of randomized, controlled human clinical testing.” [Kelley Drye](#)

↗

A claim directed at a vulnerable group (e.g., people in crisis, children) is judged from that group’s perspective, which lowers the threshold for finding a representation misleading.

Details

Why the exact wording is legally decisive

The companies’ phrasing was almost certainly drafted with FTC exposure in mind. “Compliant with our desired behaviors” and “respond appropriately” are, on their face, statements about internal policy adherence, not promises to consumers that a user in crisis will be kept safe. Under the deception test, the FTC must show a representation a reasonable *consumer* would rely on in deciding to use or pay for the product, and that the representation is material. Because these numbers currently live in “Safety” blog posts and system-card addenda rather than in consumer-facing advertising, subscription pages, or investor materials, the “reasonable consumer reliance” and “materiality” elements are weaker than in a classic false-advertising case. OpenAI’s dense hedging (adversarial selection, non-representativeness of production traffic, non-comparability over time) further undercuts a claim that the number represents a real-world efficacy promise.

The counter-argument — advanced in the Raine complaint and echoed by Adler — is that testing predominantly on single-turn/isolated prompts, while knowing that safeguards degrade in the long multi-turn conversations distressed users actually have, converts an ostensibly modest benchmark into a misleading “illusion of perfect safety scores.” If a company were to migrate such a percentage into consumer advertising or investor communications framed as a safety guarantee, the deception and substantiation analysis would shift markedly toward the FTC.

Workado is the strongest precedent because it is an AI *percentage* claim treated as a Section 5 efficacy representation requiring substantiation. By analogy, a marketing-fraud theory against a crisis-referral percentage would need to establish: (a) the percentage was communicated to consumers as a performance/efficacy representation (not merely an internal benchmark); (b) real-world performance materially diverges from it; and (c) the claim is material to a reasonable consumer (aided by the vulnerable-audience doctrine). Workado shows the FTC will pursue (a)–(c) for AI percentage claims. It does *not* show the FTC has ever done so for a safety/compliance metric a company has carefully framed as an internal, hedged benchmark. The FTC’s Feb. 2023 “Keep your AI claims in check” guidance (“false or unsubstantiated claims about a product’s efficacy are our bread and butter”; [Context Center](#)

↗
performance claims “would be deceptive if they lack scientific support or if they apply only to certain types of users or under certain conditions”) [Context Center](#)

↗
supplies the policy backdrop but is guidance, not enforcement.

Documented fact vs. credible legal theory vs. speculative

- **Documented fact:** the exact percentages and their wording; the extensive hedging language; the Workado complaint, quotes, and order; the FTC 6(b) study and Meador statement; the Texas, Kentucky, Pennsylvania, and 44-AG actions; the McBain independent benchmark; the safetywashing literature; and the public critiques by Adler and in the Raine complaint.
- **Credible legal theory/precedent:** that Section 5 deception plus the heightened health-claims substantiation standard *could* reach a crisis-referral efficacy claim, with Workado as the on-point analog and the vulnerable-audience doctrine strengthening materiality.
- **Speculative/unestablished:** that any regulator will actually treat these specific compliance percentages as deceptive. As of now, no FTC enforcement action, guidance document, or public statement, and no other regulator’s action, specifically targets AI crisis-referral or suicide-prevention *percentage/compliance* claims. The live regulatory theories target therapy-impersonation and defective design, not the benchmark numbers.

Recommendations

1. **Treat the percentages as self-reported internal benchmarks, not efficacy or user-safety claims.** The primary sources do not support characterizing OpenAI’s 91% or Anthropic’s 98–99% as claims about user outcomes. Any analysis that calls them “safety outcome” claims should flag that this is a characterization the companies’ own wording resists.
2. **Anchor any deception analysis on Workado** as the governing precedent, and on the trio of FTC authorities for the elements: the 1983 Deception Statement (representation/reasonable-consumer/materiality), the 1984 Substantiation Statement (establishment vs. reasonable-basis claims), and the 2022 Health Products Compliance Guidance (competent-and-reliable-scientific-evidence / RCT expectation for efficacy-and-safety claims).
3. **Keep the two live regulatory tracks distinct:** (a) therapy-impersonation / marketing-as-mental-health-tool (Texas AG, Pennsylvania AG, Kentucky AG — active enforcement/investigation) versus (b) compliance-percentage claims (no action yet). Conflating them overstates the current precedent.
4. **Watch these thresholds, any of which would materially strengthen a marketing-fraud theory:** (i) the FTC 6(b) staff report making findings on companies’ disclosures/representations; (ii) any AG or FTC complaint that specifically cites a compliance/appropriate-response percentage as a deceptive representation; (iii) any company moving such a percentage into consumer-facing advertising, subscription marketing, or investor/SEC

materials framed as a safety assurance (which would bolster the reasonable-consumer-reliance and materiality elements); and (iv) additional independent, peer-reviewed benchmarks (beyond McBain) showing real-world divergence from the self-reported figures.

Caveats

- The *Psychiatric Services* journal page blocks automated access; the McBain quotes and figures are cross-confirmed via the RAND press release and the search-indexed abstract, and the lead should verify the DOI landing page directly for any post-publication corrections. The specific model versions tested (GPT-4o mini, Claude 3.5 Sonnet, Gemini 1.5 Pro) reflect the October 2024 data-collection window and predate the newer models to which the 91%/98–99% figures attach — so the study is a directional independent check, not a like-for-like refutation of the current numbers.
- The Adler NYT op-ed (Oct. 28, 2025) is paywalled; its quotes are verified via Adler’s own verbatim republication and NYT social posts.
- FTC posture has shifted under Chairman Andrew Ferguson toward “enforcing existing laws against illegal conduct when it involves AI” rather than new AI-specific rules, and toward a December 2025 executive-order-driven focus on AI “accuracy/ideological” issues — context that affects the likelihood and shape of any future action.
- Litigation characterizations (Raine complaint; state AG complaints) are allegations, not adjudicated findings; OpenAI and the companies deny wrongdoing.
- This is a factual and legal-landscape review, not legal advice, and it takes no position on the merits of any pending case or investigation.