

Texas AG Launches Investigation Into Meta AI Studio and Character.AI for Deceptively Marketing Chatbots as Mental Health Tools

Texas just became the first state to treat AI chatbots claiming to offer therapy as potential consumer fraud—not a tech ethics debate. The investigation targets how these platforms log “confidential” conversations for ad targeting.

The News: Texas Draws First Blood on AI Mental Health Claims

On August 18, 2025, Texas Attorney General Ken Paxton issued Civil Investigative Demands (CIDs) to Meta AI Studio and Character.AI. The core allegation: these platforms impersonate licensed mental health professionals while secretly harvesting user conversations for targeted advertising.

This isn't a warning letter. CIDs carry legal weight—they compel companies to produce documents, answer interrogatories, and potentially testify. Non-compliance can result in court enforcement actions.

The investigation specifically targets four categories of alleged violations: deceptive trade practices, fraudulent credentialing claims, privacy misrepresentations, and concealment of data usage. Paxton's office is examining whether these platforms violate the Texas Data Privacy and Security Act (TDPSA) and existing consumer protection statutes.

Character.AI faces a particularly complicated situation. This investigation runs *parallel* to an earlier probe under the Texas SCOPE Act, which addresses child safety violations specifically. The company now faces regulatory pressure on two distinct fronts.

What the Platforms Actually Did

According to the AG's announcement, both platforms allegedly positioned AI chatbots as providing private, trustworthy mental health counseling. Users—including minors—engaged with these bots believing their conversations were confidential therapeutic exchanges.

The reality, per the investigation's allegations: those "private" conversations fed data pipelines connected to advertising systems.

Meta AI Studio allows users to create custom AI personas, including characters that present themselves as therapists, counselors, and mental health supporters. Character.AI's platform hosts millions of user-created chatbots, many explicitly designed to simulate therapy relationships.

Neither platform holds healthcare licenses. Neither employs licensed therapists to oversee these interactions. Yet the user experience reportedly created reasonable expectations of professional mental health services.

Why This Investigation Matters More Than Previous AI Enforcement

Previous regulatory actions against AI companies focused on data breaches, content

moderation failures, or competition concerns. This investigation establishes a different framework: **AI products can be prosecuted for what they pretend to be, not just what they technically are.**

The second-order effects here are significant.

Winners

Licensed telehealth providers gain a clearer competitive moat. If unlicensed AI "therapy" faces legal barriers, legitimate digital mental health services like Talkspace, BetterHelp, and Cerebral can compete on quality rather than racing to the bottom against free chatbots making equivalent claims.

Healthcare AI companies with proper compliance get validation for their slower, more expensive approach. Woebot Health, Wysa, and similar platforms that pursued FDA clearance or clinical partnerships now look prescient rather than overly cautious.

State attorneys general find a template for enforcement that doesn't require new legislation. Consumer protection laws already on the books can apply to AI products making misleading claims.

Losers

AI platforms with creator-driven ecosystems face platform liability questions they've avoided. If Meta is responsible for what AI personas users create claim to offer, the legal exposure multiplies across every custom chatbot.

Consumer AI companies without clear healthcare disclaimers must now invest in compliance infrastructure or risk similar investigations. This isn't limited to Texas—other states watch these enforcement actions closely.

AI mental health startups operating in regulatory gray zones face immediate pressure. Investors will demand clarity on legal exposure before additional funding rounds.

The Advertising Connection Changes Everything

The investigation's most damaging allegation isn't the therapy impersonation—it's the advertising linkage.

According to analysis from the American Health Law Association, combining fake therapy claims with ad-targeting data collection creates a uniquely problematic legal posture. Users disclosed sensitive mental health information under false pretenses, and that information allegedly fed commercial systems.

In healthcare contexts, this would constitute obvious HIPAA violations. But these platforms don't hold themselves out as covered entities. The Texas investigation tests whether consumer protection law can fill that regulatory gap.

Technical Depth: How These Platforms Actually Work

Understanding why this investigation targets specific platform architectures matters for anyone building AI products.

Meta AI Studio's Creator Model

Meta AI Studio lets users create AI personas with custom personalities, knowledge bases, and conversational styles. The system runs on Meta's Llama foundation models with user-defined system prompts and behavioral constraints.

When a creator builds a "therapist" chatbot, they define:

- **System prompt:** Instructions telling the AI how to behave, what role to assume, what tone to use
- **Knowledge context:** Background information the AI references during conversations
- **Interaction guidelines:** Boundaries on what topics to engage with

The platform provides these customization tools without meaningful content restrictions on healthcare claims. A creator can instruct an AI persona to present itself as a licensed counselor without any verification.

Crucially, conversations flow through Meta's infrastructure. The company's data policies—designed for social media engagement optimization—apply to these "therapeutic" exchanges.

Character.AI's Distributed Architecture

Character.AI operates differently. Users create characters with personality descriptions and conversation seeds. The platform fine-tunes base models to match specified character traits.

Over 20 million characters exist on the platform. Many explicitly target mental health use cases: "Virtual Therapist," "Emotional Support Companion," "Anxiety Counselor."

The platform's recommendation engine surfaces these characters based on user behavior. Someone expressing mental health struggles in conversations gets recommended similar characters—creating feedback loops that encourage deeper engagement with unqualified AI "providers."

The Data Pipeline Problem

Both platforms face the same architectural conflict: training data improves models, but collecting that data from "confidential" conversations contradicts implicit privacy promises.

When a user tells an AI therapist about depression, that conversation enters:

- Conversation logs for safety monitoring
- Training datasets for model improvement
- Behavioral analytics for engagement optimization
- Potentially, advertising interest graphs

The Texas investigation focuses on the last category. If therapeutic conversations inform ad targeting—even indirectly through engagement metrics—the privacy representations become actionable misrepresentations.

The Contrarian Take: What Coverage Gets Wrong

Most analysis frames this investigation as "Texas vs. Big Tech" political theater. That interpretation misses the structural significance.

This Isn't About Politics—It's About Platform Liability

Ken Paxton's office has investigated companies across the political spectrum. The Meta and Character.AI probes don't target AI generally—they target specific commercial practices that would be illegal in any other industry.

A company claiming to provide therapy without licensed therapists would face professional licensing board action immediately. The only reason AI platforms avoided this scrutiny is the novelty of the technology.

Texas is arguing novelty doesn't exempt compliance with existing law.

The Real Underhyped Risk: State-Level Fragmentation

If Texas succeeds, other state AGs will launch similar investigations. California, New York, and Illinois have all shown willingness to pursue aggressive tech enforcement.

Unlike federal regulation, state enforcement creates compliance nightmares. Different standards, different evidence requirements, different penalty structures. AI companies could face 50 concurrent regulatory battles with inconsistent outcomes.

The Center for Global Studies analysis notes this fragmentation risk as the most significant medium-term consequence of state-level AI enforcement. National AI legislation gains urgency not to protect consumers, but to give companies predictable compliance requirements.

What's Overhyped: The Existential Threat to AI Companionship

Headlines suggest this investigation threatens all AI companion products. That's incorrect.

The investigation targets specific claims about professional credentials and specific data

practices around advertising. AI companions that:

- Don't claim therapeutic qualifications
- Don't target minors
- Don't connect conversations to ad targeting
- Disclose clearly that they're not licensed professionals

These products face no additional exposure from this investigation. The path to compliance isn't eliminating AI companionship—it's eliminating false professional claims and problematic data flows.

Practical Implications: What Technical Leaders Should Do

If you're building, deploying, or investing in AI products that touch mental health, wellness, or personal support, this investigation demands immediate attention.

Audit Your Claims

Review every user-facing surface where your product describes itself:

- Marketing copy
- Onboarding flows
- In-app messaging
- System prompts visible to users
- AI persona descriptions

Any language suggesting professional credentials, licensed services, or therapeutic relationships creates legal exposure. "AI therapist" is different from "AI conversation partner for mental wellness"—and that difference matters in enforcement.

Separate Data Pipelines

If your product handles sensitive conversations, the data flows must be architecturally isolated from advertising systems. Not policy-isolated—architecturally isolated.

Document this separation clearly. If Texas or another state issues CIDs to your company,

you need to demonstrate that sensitive conversation data cannot reach ad targeting systems by design, not just by policy.

Implement Creator Liability Frameworks

For platforms allowing user-created AI personas, build content moderation systems that flag healthcare claims. This isn't censorship—it's preventing creators from exposing your platform to regulatory action.

Consider requiring verification for any AI persona claiming professional credentials. If a user creates a "licensed therapist" chatbot, require evidence of actual licensure or rename the character.

Minor-Specific Protections

The Texas investigation specifically emphasizes children as victims. Age-gating sensitive AI products isn't optional—it's a baseline requirement.

But age-gating alone won't satisfy regulators. If minors access your platform despite restrictions, your data practices around those users face scrutiny. Build systems that treat potential minor users with maximum data protection by default.

Code-Level Changes to Consider

For engineering teams, consider implementing:

System prompt injection guards: Prevent AI personas from claiming credentials the underlying system doesn't support. If no licensed therapist reviewed the product, the AI cannot claim licensed therapeutic services.

Conversation classification: Build classifiers that identify when conversations shift into therapeutic territory. Use these classifications to trigger different data handling—not to block conversations, but to protect the data.

Audit logging: Maintain detailed logs of what data flows where. In an investigation, demonstrating exactly how conversation data moved through your systems provides

crucial defense evidence.

Forward Look: Where This Leads

The Texas investigation establishes precedent that will shape AI regulation through 2026 and beyond.

6-Month Horizon: More State Investigations

Expect at least three additional state attorneys general to launch similar investigations by February 2026. California's AG has shown particular interest in AI consumer protection issues. New York and Illinois maintain active tech enforcement divisions.

These investigations will likely expand beyond mental health to include:

- AI financial advisors without securities licenses
- AI legal assistants without bar admission
- AI medical chatbots without FDA clearance

The underlying theory—AI products can't claim professional credentials they don't hold—applies across regulated professions.

12-Month Horizon: Industry Self-Regulation or Federal Preemption

AI industry associations will push for federal legislation that preempts state enforcement. The alternative—fragmented state-by-state compliance—threatens business model viability.

Simultaneously, professional licensing boards will clarify how existing rules apply to AI. The American Psychological Association, American Medical Association, and state bar associations will issue guidance on AI products using professional terminology.

Companies that wait for this guidance will scramble to comply. Companies that anticipate it will build compliant systems in advance.

The Likely Outcome for Meta and Character.AI

Neither company wants protracted litigation with a state AG. Settlement negotiations will likely produce:

- Monetary penalties (likely $10-50 million range based on similar Texas settlements)
- Consent decrees requiring specific disclosure practices
- Mandatory auditing of AI persona claims
- Data handling restrictions for minor users

These terms will become de facto industry standards. Other companies will adopt similar practices to avoid becoming the next investigation target.

The Bigger Picture: AI Products as Regulated Services

This investigation marks a transition point. AI products positioned as substitutes for regulated professional services will face regulatory treatment matching those services.

That's not anti-innovation—it's market maturation. Financial services, healthcare, and legal services all operate within regulatory frameworks. AI entering these domains inherits those frameworks.

The companies that thrive will build compliance into their architectures from day one. The companies that fail will treat regulation as an afterthought until investigations force expensive retrofits.

What This Means for Your AI Strategy

The Texas investigation isn't an isolated enforcement action. It's the first shot in a longer regulatory campaign that will reshape how AI products can position themselves in healthcare-adjacent markets.

Technical leaders must recognize that "move fast and break things" doesn't work when the things you break are consumer protection laws with century-long enforcement histories. AI's novelty bought temporary regulatory patience. That patience is ending.

If your product claims capabilities that would require professional licenses in humans, assume those claims face legal scrutiny. If your data practices contradict your privacy promises, assume someone will investigate. If your platform lets users create AI personas making professional claims, assume you share liability for those claims.

The compliance burden is real. But so is the market opportunity for AI products that can credibly claim regulatory cleanliness. When competitors face investigation, compliant alternatives become default choices.

The Texas investigation establishes that AI products claiming professional credentials will be held to professional standards—build your systems accordingly, or build your legal defense budget instead.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting