quinn emanuel
quinn emanuel urquhart & sullivan, llp

# Defamation in the AI Era

*New technologies breed new lawsuits. Social media did it. Generative AI is doing it now.*

ChatGPT, Claude, and other large language models produce statements that sound authoritative—even when fabricated. LLMs can generate false statements about real people, including invented lawsuits, criminal records, or accusations. They often do so in a tone that seems authoritative. Predictably, such outputs have led to defamation lawsuits.

Early results favor AI developers. In May 2025, a Georgia state court granted summary judgment to OpenAI on a claim based on ChatGPT output that was indisputably false. The court reasoned that the user knew ChatGPT might fabricate, so a reasonable reader in the user's position would not have understood the output as a statement of fact.

The decision reassures developers—but on narrow grounds. Harder cases involving different facts are also pending: users unaware of hallucination risk, false statements republished to millions, or companies that ignore repeated warnings. Here are some aspects of AI-based defamation liability worth knowing.

## I. How AI Creates Defamatory Content

LLMs are designed to generate text—not retrieve it—by predicting the next word in a sequence based on patterns learned during training. When reliable information is missing, the model fills the gap. Sometimes it does that by fabricating facts, sources, or entire narratives. Courts are now encountering several common patterns.

*Mixing True Facts to Create False Statements:*

AI can state facts that are individually accurate but combine them into something false and harmful. *Battle v. Microsoft Corp.*, No. 1:23-cv-01822 (D. Md. filed July 7, 2023), illustrates the problem. Bing's AI summary conflated two people: Jeffery Battle, an aerospace educator, and

Jeffrey Battle, a convicted terrorist. The output read:

> Jeffrey Battle, also known as The Aerospace Professor, .... ***However***, ***Battle*** was sentenced to eighteen years in prison after pleading guilty to seditious conspiracy and levying war against the United States. He had two years added to his sentence for refusing to testify before a grand jury.

The word "however" fused two biographies into one defamatory portrait. This case may not yield published guidance: On October 23, 2024, the court compelled arbitration.

*Implying False Facts*:

A statement need not be literally false to defame. It can imply harmful facts through context or false attribution. In November 2024, Indian news agency ANI sued OpenAI in the Delhi High Court, alleging that ChatGPT generated fabricated content that was falsely attributed to ANI. This included a purported interview/podcast appearance by a political figure that never occurred. The case also raises copyright-related claims. In 2025, industry stakeholders, including T-Series, Saregama, Sony Music, and publisher associations, applied to intervene or join the case. Given the novel issues raised, the court appointed two amici curiae, including to assist on threshold questions such as jurisdiction, which OpenAI has contested. The matter remains pending.

In the U.S., related claims are emerging under trademark law. *The New York Times*, *The Wall Street Journal*, and others have asserted trademark dilution claims against AI platforms, alleging that false attribution of AI-generated content to their publications harms their brands. As image and video generation tools mature, similar claims may arise from visual or audiovisual fabrications.

In March 2025, Minnesota solar company Wolf River Electric sued Google after its AI Overview falsely stated that the Minnesota Attorney General had sued the company for deceptive practices. The complaint alleges that the Google tool fabricated the statement and that it was not supported by the cited sources. In the wake of its publication, customers cancelled contracts. The company seeks over $100 million in damages. *See LTL LED, LLC v. Google LLC*, Case No. 25-cv-02394 (D. Minn.), Dkt. No. 1-1 (Compl.). In January 2026, the court remanded the case to Minnesota state court.

Image-generation tools are creating similar issues. In January 2026, a plaintiff filed a class action against xAI, alleging that its Grok chatbot generated sexualized deepfake images based on photos she had posted to X. The complaint asserts defamation claims on the theory that the images "gave the false impression that Plaintiff and Class Members were actually photographed in a revealing or sexualized manner." *Doe v. xAI Corp.*, No. 5:26-cv-00772 (N.D. Cal. filed Jan. 23, 2026). The case is in its earliest stages, but it signals that defamation-by-implication claims may extend beyond text to AI-generated images and video.

*Pure Fabrication*:

Sometimes LLMs invent facts from nothing. In *Walters v. OpenAI*, Case No. 23-A-04860-2 (Ga. Super. Ct. Gwinnett County, filed July 17, 2023) , a journalist asked ChatGPT to summarize a complaint involving the Second Amendment Foundation. ChatGPT responded that radio host Mark Walters was a defendant accused of embezzling from the organization. The lawsuit did not exist. The accusation was invented. When prompted further, ChatGPT elaborated on the fabricated claims. In May 2025, the court granted summary judgment to OpenAI, holding that—under the reasonable-reader standard, which allows "time for reflection"—no reasonable reader in the journalist's position could have understood ChatGPT as communicating "actual facts." *Walters v. OpenAI*, No. 23-A-04860-2, Order at 7 (Ga. Super. Ct. Gwinnett Cty. May 19, 2025).

The court emphasized the surrounding context: ChatGPT warned it could not access the linked complaint and noted that the relevant lawsuit post-dated its "knowledge cutoff date," and OpenAI repeatedly warned (including in its Terms of Use) that ChatGPT "may produce inaccurate information." *Id.* at 6. The journalist was also familiar with ChatGPT's "flat-out fictional responses," had the actual complaint available to verify the output, and testified that within about ninety minutes he determined the output "was not true." *Id.* at 6–7.

## II. Defamation Basics Still Apply to AI-based Claims

Showing that an LLM fabricated harmful, false statements is not enough to sustain a defamation claim. The plaintiff must establish: (1) publication to a third party, (2) of a false statement of fact about the plaintiff, (3) made with the requisite fault, (4) that harmed the plaintiff's reputation.

*Publication and Defamatory Meaning:*

Publication does not require a mass audience. If the AI platform (or its owner) is deemed the speaker—rather than the user who entered the prompts—the element may be satisfied when the AI platform delivers its response to even a single user.

The harder question is whether AI outputs can convey facts, not fiction. In *Walters v. OpenAI*, the Georgia court held they did not, under the circumstances present. The court emphasized four facts: (a) ChatGPT warned the journalist that it could not access the document described in his prompt; (b) OpenAI's terms of use stated that outputs may contain inaccuracies; (c) the journalist had a copy of the actual complaint to verify; and (d) the journalist quickly recognized the output as fabricated.

But *Walters* leaves open the question of what happens when users *do not* know the statements are hallucinations. Two pending cases test that scenario.

In *LTL LED, LLC v. Google LLC* (Minn.), a solar company alleges that Google's AI Overview falsely told the public that it faced a Minnesota Attorney General lawsuit for deceptive practices. Unlike *Walters*, the AI-generated statements reached the general public—not a savvy journalist with independent knowledge—and appeared alongside cited sources that lent an air of reliability. In response, customers canceled contracts. The case is still in its pre-trial phase.

Similarly, in *Starbuck v. Google LLC*, No. N25C-10-211 MAA (Del. Super. Ct., filed Oct. 2025), a conservative activist alleges that Google's Bard and Gemini chatbots generated fabricated statements—including sexual assault accusations, criminal records, and invented court documents—and attributed them to fictitious sources. The complaint details how third parties, including a "mom's group" evaluating whether to support Starbuck's business initiatives, relied on the AI-generated content. Google has moved to dismiss. It argues, among other things, that no identifiable audience relied on the outputs and that the AI tools explicitly warned users of possible inaccuracies. The case remains pending.

Courts will ultimately decide whether growing public awareness of hallucinations—or companies' express disclaimers—limits whether AI outputs can reasonably be understood as asserting defamatory facts. That conclusion could significantly limit defamation liability for AI-generated statements. Courts may not go that far, given that AI tools are increasingly marketed as powerful, accurate research assistants.

*Falsity*:

Truth is among the most powerful defamation defenses. But in most AI defamation cases, falsity is undisputed.

*Fault:*

Establishing fault presents the largest obstacle for AI defamation plaintiffs.

*Public Figures and Actual Malice:*

A public figure must prove actual malice: The defendant subjectively knew that the statement was false or entertained serious doubts about its truth. These standards were built for human speakers. It seems implausible to say that a probabilistic language model "knew" whether its next predicted word was true, let alone "entertained serious doubts." And how can the AI platform, which controls neither the user's prompt nor the spontaneous output, possess the requisite knowledge about statements it did not know the model would generate?

*Walters* addressed this head-on. It rejected the argument that merely releasing software capable of hallucinations establishes fault under either the actual malice or negligence standards. OpenAI introduced expert testimony describing its mitigation efforts to reduce hallucinations. The court found no evidence that OpenAI knew the specific output was false or acted with reckless disregard.

*Private Figures and Negligence:*

Private-figure plaintiffs generally face a lower fault standard—negligence—but they still must show the defendant failed to exercise reasonable care under the circumstances. In defamation cases, courts often evaluate what reasonable care would require in the defendant's position. This can include the availability and use of safeguards to prevent false statements. In the generative AI context, this suggests developers may seek to defeat negligence claims by demonstrating robust, industry-consistent mitigation measures. In *Walters*, the court credited evidence of OpenAI's hallucination-reduction efforts in concluding the plaintiff had not raised a triable negligence claim.

*The Role of Notice.*

The fault question may turn on what happens *after* plaintiffs notify AI companies of false outputs. Some argue that a company's refusal to correct flagged falsehoods could establish knowledge. Under traditional defamation principles, however, fault is typically measured at the time of publication, not afterward.

*Starbuck v. Meta* tested this theory. The plaintiff alleged that he repeatedly notified Meta of fabricated January 6th accusations and Holocaust denial claims, yet Meta failed to correct them. The case settled in August 2025. Reports indicated that, as part of the settlement, Meta agreed to hire the plaintiff as a consultant to address political bias in its AI. Although the settlement terms are undisclosed, the case illustrates how pre-suit notice and a defendant's failure to respond may strengthen a plaintiff's position—or at least provide leverage.

*Starbuck v. Google* presents similar facts: The plaintiff alleges that he notified Google of the defamatory outputs several times over two years. Yet according to the complaint, Google "sat back and did nothing." Whether courts credit this theory of post-notice liability remains to be seen. Google's motion to dismiss is pending.

*Harm*.

Damages for reputational harm can be difficult to quantify. Plaintiffs may recover presumed damages in certain circumstances, but this remedy may require a showing of actual malice where the speech involves a matter of public concern—even for private figures. Given the difficulty of proving actual malice in the AI context, most plaintiffs will be limited to actual damages proximately caused by the defamatory statement.

This limitation may discourage some claims—but not all. In *LTL LED*, the plaintiff alleges documented contract cancellations totaling hundreds of thousands of dollars, with claimed damages ranging from $110 million to $210 million. Where AI-generated falsehoods cause concrete, provable business losses, the damages calculus changes significantly.

By contrast, Walters testified that he had not suffered actual damages, and the court concluded that he could not recover presumed or punitive damages on the record presented—resulting in summary judgment for OpenAI.

## III. Takeaways

Beyond the doctrinal questions discussed above, courts will also need to address immunity. Section 230 of the Communications Decency Act shields website operators from liability for content "provided by another information content provider." How Section 230 applies to AI-generated outputs—particularly where the output is arguably the company's own generated speech rather than third-party content—remains an unsettled question.

*For Potential Plaintiffs:*

AI defamation claims present unique challenges, but also distinct advantages.

*First*, falsity may be easier to establish. Hallucinated outputs are often demonstrably fabricated, narrowing the dispute and reducing the likelihood the case turns on proving truth.

*Second*, fault may be easier to establish against *users* who republish AI-generated falsehoods. Unlike traditional defamation cases, there will be a contemporaneous record: the prompts entered, the outputs received, and what the user did next. That record can reveal whether the user knowingly published statements whose truth they had reason to doubt.

*Third*, plaintiffs should notify AI companies promptly and request correction. Pre-suit notice creates a record. If the company ignores it, that may support a fault argument—or at least provide settlement leverage, as in *Starbuck v. Meta*.

*For AI Platforms:*

Implement a protocol for responding to complaints about false outputs. Ignoring notice is a bad fact.

Design systems to warn users when outputs may be unreliable, particularly when the model lacks access to the information requested. The *Walters* judge credited OpenAI's warnings and other contextual facts in concluding that a reasonable reader in that setting would not have understood the output as stating fact.

Continue educating courts on how LLMs work. These models generally generate text by

predicting sequences rather than retrieving documents like a search engine (though outputs may sometimes resemble or reproduce training material). That distinction matters for how courts apply traditional defamation principles to a new context.

*For Companies and Individuals Using AI Tools:*

Liability is not limited to AI developers. Anyone who republishes AI-generated falsehoods may face defamation claims.

Before publishing any AI-generated factual assertion, verify it independently and document that process. Use enterprise versions of AI tools where available—they typically include stronger safeguards. These steps may help defeat a negligence claim if a falsehood slips through.

***

If you have any questions about the issues addressed in this memorandum, or if you would like a copy of any of the materials mentioned in it, please do not hesitate to reach out to:


Bobby Schwartz
Partner
Los Angeles
robertschwartz@quinnemanuel.com, Tel: (213) 443-3675

Mari Henderson
Partner
Los Angeles
marihenderson@quinnemanuel.com, Tel: (213) 443-3364

Marie Hayrapetian
Associate
Los Angeles
mariehayrapetian@quinnemanuel.com, Tel: (213) 443-3552


To view more memoranda, please visit www.quinnemanuel.com/the-firm/publications/
To update information or unsubscribe, please email updates@quinnemanuel.com