LOS ANGELES & SAN FRANCISCO

Daily Journal

www.dailyjournal.com

FRIDAY, FEBRUARY 6, 2026

Defamation by chatbot: Why Section 230 doesn't protect the new tech

For 30 years, Section 230 insulated platforms from liability, but generative AI is forcing courts to ask a new question: When AI speaks, is it third-party content, or the provider's own?

By Krista L. Baughman

Shutterstoc

For 30 years, technology companies have operated with assurance that Section 230 of the Communications Decency Act broadly shielded them from liability tied to third-party content. But with the advent of novel artificial intelligence technology, Section 230's protections are now being tested by plaintiffs who allege harm at the hands of generative AI. One example is the pending case of *Wolf River Electric v. Google LLC*, in which a Minnesota solar company alleges that Google's AI Overview falsely told its potential customers that it had been sued by the state attorney general for deceptive practices.

As courts begin to confront the application of Section 230 to this new frontier, they will need to determine whether generative AI responses to user queries constitute third-party content–or instead, whether AI providers are themselves responsible for what their systems generate.

Section 230's text and limitations

Congress enacted Section 230 in 1996 to promote free expression on the then-burgeoning internet while protecting platforms from the burdens of policing vast amounts of third-party content. The first policy objective–promotion of free expression–is reflected in subsection (c)(1)'s immunity for a platform that passively hosts "information provided by another information content provider." This codifies Congress's choice to protect companies that "serve as intermediaries for other parties' potentially injurious messages," as recognized by the 9th Circuit in *Zeran v. AOL*. The second objective–encouraging ISPs to voluntarily block or filter offensive material without fear of liability–is reflected in subsection (c)(2)'s protection of "good faith" efforts to moderate content.

These are the primary protections Section 230 provides. But when evaluating the statute's scope, it is equally important to consider what the statute does not do–including protect ISPs for their own unlawful speech.

This was established in the seminal case of *Fair Housing Council v. Roommates.com*, which held that platforms lose protection when they "materially contribute" to content creation. The 9th Circuit held that by requiring users to disclose protected characteristics in a drop-down menu as a condition of accessing its services, Roommates.com took an active role in creating discriminatory content, thus stripping immunity.

Since 1996, courts have dramatically expanded the scope of Section 230, immunizing platforms from all manner of legal claims, including those involving user posts, terrorist recruitment materials, false dating profiles and use of algorithms to present third-party content. But notably, in each of these examples, the underlying speech at issue was generated by third parties. This presents a legally significant difference from cases alleging defamation by generative AI.

Defamation-by-chatbot falls outside Section 230

As of today, no court has explicitly extended Section 230 immunity to generative-AI outputs. And there are compelling textual, legal and policy reasons why this should remain the status quo.

Based on its text, Section 230 addresses liability for information generated by third parties–not for new outputs generated by AI companies. In *Wolf River*, for example, Google's AI Overview is alleged to have generated novel assertions about plaintiff company that existed nowhere on the internet, and to have cited to "sources" that did not contain the defamatory assertion the AI had published. Factually, this is a marked departure from the functioning of traditional search features, which display links and excerpts from the websites that first published the content.

Indeed, the creative capacity of artificial intelligence is why we call it "generative" AI: Its value lies in

creating fresh content, not regurgitating existing text. If ChatGPT simply copied and pasted existing speech from other websites, the output would be redundant of search browser functionality and unhelpful to users.

While AI companies cannot deny that their products do more than copy and paste from other parts of the internet, they may argue that chatbots "remix" information found in the training data, thereby justifying Section 230 immunity. This argument should fail in cases where a plaintiff can prove that the false information never existed in *any form* prior to being published by a chatbot. Moreover, as Professor Eugene Volokh observes in *Large Libel Models? Liability for AI Output*, all human speech is a product of rearranging encountered words and patterns, yet that reality does not immunize humans from assembling defamatory statements. AI's generative nature places AI companies squarely in the role of content creator, not neutral platform for third-party speech. For these reasons, subsection (c)(1) cannot protect generative-AI outputs.

Section 230(c)(2) is equally unlikely to provide protection, as the text assumes platforms moderate someone else's content and thus would not immunize outputs that are created by a platform's own generative systems. Moreover, (c)(2) immunity is contingent on the provider acting in "good faith," so allegations of bad faith–such as libel–would undermine its applicability.

Finally, from a legal and policy perspective, granting blanket immunity for AI platforms should give considerable pause, as it would create an accountability vacuum that has never before existed in defamation law. In traditional cases of online libel, while the victim may be prevented by Section 230 from suing the platform directly, she can (at least, in theory) pursue a claim against the original, human publisher of the speech. However, no third-party human speaker exists in AI defamation cases, thus precluding a victim from seeking relief. This outcome is particularly worrisome given the significant institutional credibility carried by AI outputs: When the imprimatur of Google, OpenAI or Meta appears above fabricated accusations, readers are likely to trust them. Depriving individuals of redress for the destruction of their good names and reputations would be an unprecedented result that should be given serious consideration.

A different technology demands a different approach

Recently, calls for AI regulation have intensified, yet no federal legislation has emerged. In this vacuum, litigants will continue to assert defamation liability against AI companies for generative outputs, and courts must contend with whether a law designed for passive intermediaries can govern generative AI. Fortunately for the courts, Section 230 need not be rewritten in this process–because the statute never promised immunity for original content in the first place.

Krista L. Baughman *is a trial attorney and litigator of First Amendment and related civil liberties issues. She is the founder of Baughman Law PC.*