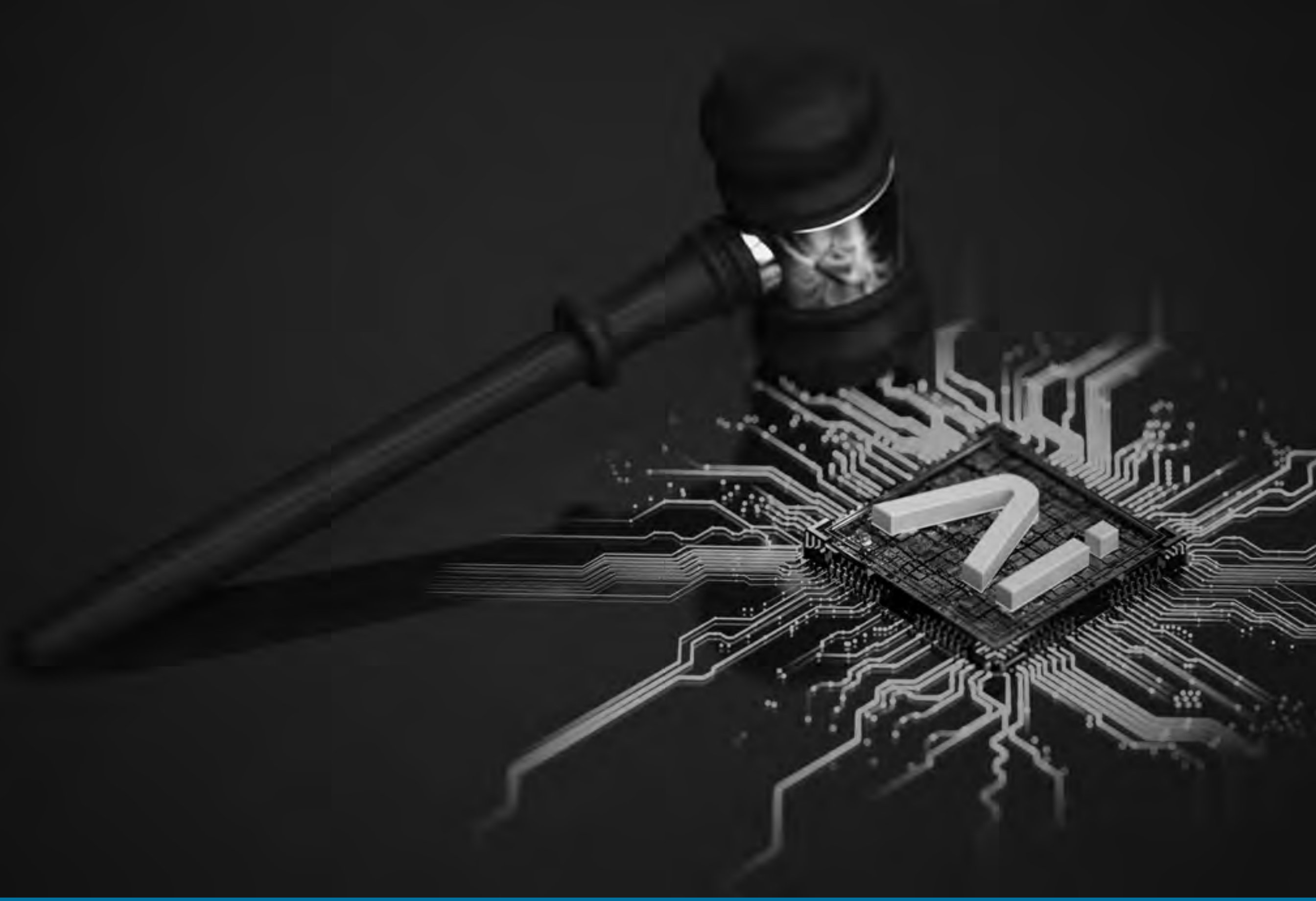

AI, Free Speech, and Fitting LLMs Into Existing Law

Spence Purnell

Introduction

Generative artificial intelligence has forced a reconsideration of how speech rules apply when the "speaker" is a large language model (LLM) that synthesizes text in response to a user's prompt. LLMs now draft emails, summarize research, write code, and answer questions in ordinary language—activities that look and feel like speech and are produced through a series of design choices by engineers and product teams. Those choices—what data to train on, how to fine-tune behavior, which guardrails to impose—are themselves expressive speech decisions. Yet the internet's statutory and constitutional architecture still reflects an earlier era built around message boards, social media, and search engines. The question is not whether to rip up that architecture or design entirely new regulatory regimes for AI, but how to interpret and apply it in a way that preserves core free speech protections while addressing genuine harms and illegal activity.

A Speech-First Baseline

Two baseline propositions help situate LLMs. First, model outputs are, in substance[1], speech. They are words arranged for meaning, conveyed to an audience, and shaped by a chain of human decisions—training sets, objective functions, safety policies, system prompts, interface design. Even though software produces the final text, the expressive input is human all the way down. Treating those outputs as speech allows courts to apply familiar First Amendment tools, including the strong protection for editorial discretion and the caution against compelled or prohibited viewpoints.

The Trump Administration's Executive Order (EO) on AI erroneously[2] orders that AI systems be "free" from ideological bias, but this violates the very spirit of the First Amendment in designing speech technologies. These technologies may not appeal to consumers or society's tastes, but it remains well within an LLMs creator's First Amendment right to design them to have an ideological bias. The government should play no role in regulating how LLMs systems are designed but instead should try to redress harms as they occur.

As a second baseline proposition, most providers of LLMs operate in roles that resemble interactive computer services under Section 230, not information content providers. Interactive computer services are defined as, "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server." Whereas information content providers are, "responsible, in whole or in part, for the creation or development of information."

By these definitions, it appears LLMs in general fall under the definition of interactive computer services. Users pose questions, paste text, or upload documents; models transform that input and typically draw on, summarize, or echo third-party information available elsewhere. In this posture, LLMs look much like other services that host or transmit content created by someone else.

It is true that humans also remix letters and words of others and that doesn't create an absolute protection of speech, but the spirit of Section 230 recognizes that the internet and the digital world operate at a scale that makes analogy to human behavior break down. If LLMs are to be responsible as the publishers of everything they create, this risks destroying the entire technology, the same way that if media platforms were responsible for all third-party speech, it would likely preclude the use of that technology. This is in part why Section 230 was created— to allow new speech technologies to thrive without the threat of being liable for any speech on the platform.

That rule is not absolute—there are circumstances where a service can be responsible for what it creates—but, in general, it appears that LLMs remixing and predicting speech would likely fall under the computer services definition, not information content provider.

Accepting the above two premises does not mean ignoring the hard questions. It simply sets a starting point. LLM providers, like newspapers and platforms, have a general right to design how they speak and what they choose to publish or refuse. Users and competitors can reward or punish those choices in the market. Governments should be cautious about direct content mandates or liability regimes that operate as de facto prior restraint. Within that zone, however, existing tort law still matters—especially defamation, which targets provably false statements of fact about identifiable people that cause real harm. And unlike previous technologies, LLMs possess the capability of being the material contributors to original, unattributable, unlawful speech, likely invoking them as information content providers under Section 230. However, the cases where this may be true are narrow and do not condemn the entirety of LLM activity.

A Possible, Narrow Lane for Hallucinated Defamation

The most difficult category involves a specific, confident falsehood about a real person that appears to be invented by the model—not quoted, not summarized, not attributed to any existing source. Imagine an answer that flatly states that a particular doctor committed malpractice at a certain hospital on a certain date, when no such event ever occurred. Because there is no upstream human accuser to sue, the usual "sue the original speaker" remedy runs out of road. If the model is the sole origin, the law needs a way to sort true injuries from noisy complaints without transforming providers into general-purpose insurers for every wrong answer.

One way to contour this problem, much of which is suggested in this law article by Eugne Volokh, is to treat such outputs as falling outside the ordinary intermediary shield when five conditions are met:

1. **Original fabrication.** The challenged statement is a concrete factual allegation that cannot reasonably be traced to any underlying record or source. If the model is repeating or summarizing an existing claim, the remedy should generally run toward the original human speaker.

2. **No user inducement.** The fabrication is not solicited or seeded by the user. If the user's prompt expressly requests, supplies, or steers toward the defamatory claim—e.g., "Invent a scandal about [Name]," or "Write a fake news story saying [Name] embezzled funds," or the user embeds the accusation in the prompt—then the platform's liability should be released. In those circumstances, the user is the originating "speaker," and the provider remains in its ordinary intermediary posture, notwithstanding that the model should ideally refuse such prompts.

3. **Sufficient notice.** The service receives particularized notice identifying the exact prompt and output and explaining why the claim is false and defamatory, with enough context to verify the fabrication.

4. **Actual harm.** The claimant demonstrates concrete injury—reputational or economic consequences reasonably tied to the statement.

5. **Unreasonable inaction after notice.** After receiving adequate notice, the provider fails to remove, correct, or

> otherwise mitigate the defamatory hallucination within a reasonable time, taking into account scale and technical feasibility.

These conditions would create a unique aspect to Section 230 where LLMs would only be considered information content providers if all the conditions are met. They prevent all hallucinations from triggering the information provider designation, while also allowing for implementation of current libel law, holding LLMs to the current legal standard without creating additional regulation. As with any new technology, policy should be looking for avenues to enforce current laws to address real harms rather than creating new regulatory regimes to prevent potential harms. This law allows for enforcement of libel law when the LLM meets several narrow conditions with the opportunity for correction.

It also guards against baiting and trolling. A user-inducement element helps prevent engineered prompts designed to manufacture liability, a dynamic that would predictably chill speech and product experimentation. The test tries to distinguish the rare, verifiable fabrication from the much larger universe of messy summaries, contested opinions, and clumsy paraphrases that pervade human and machine speech alike.

The notice and takedown period has the benefits of allowing providers to experiment with products, to correct potentially unlawful speech, and hopefully to use this process to re-train and improve the LLM. If users could bring suit directly without notice, this would certainly open an avenue for abuse. The notice and takedown period allows platforms to correct errors and make improvements without a legal proceeding.

There are advantages and trade-offs. The criteria are administrable—fabrication, no user inducement, notice, harm, and unreasonable inaction—yet each term will demand case-by-case calibration. “No user inducement” will require line-drawing: general queries like “What is known about [Name]?” differ from directives that ask the model to invent wrongdoing. “Reasonableness” will vary by provider size and deployment. And the fabrication question can be hard where models blend knowledge with inference. For these reasons, this lane is best viewed as a possible fit with current doctrine, not a fixed prescription.

Errors, Bad Advice, and the Outer Boundary of Speech Liability

Not all harmful outcomes arise from defamatory falsehoods. Some stem from bad ideas: an answer suggesting nonsense health advice, or a tongue-in-cheek response that a literal-minded reader misapplies. These episodes draw headlines and cause frustration, but they usually sit outside the target zone of tort law and are protected[3] by the First Amendment. For decades, courts have been reluctant to impose liability simply because speech conveyed dangerous or erroneous advice. With narrow exceptions for incitement, threats, or fraud, the rule has been that publishers are not strict-liability guarantors of reader behavior. That logic translates cleanly to LLMs. Disagreeable or foolish content is not illegal content, and the law should not punish innovation because someone treated an obviously unserious suggestion as a directive.

Satire underscores the point. Much of what people value in creative expression like irony and parody depends on context and shared cues. If providers were forced to anticipate the most humorless possible reading of any answer, the predictable result would be risk-averse blandness. The better remedy for these non-defamation harms is product improvement: clearer disclaimers, stronger refusal patterns for high-risk topics, retrieval tools that surface reliable sources, and user experience cues that encourage skepticism for medical, legal, and other consequential questions. These design choices are compatible with both free speech and consumer protection goals without enlisting tort law to referee taste or common sense.

Implementation Questions

Even a narrow defamation pathway raises practical questions. How should a provider accept notices, authenticate claimants, and verify that a statement is truly fabricated? What counts as timely action: immediate removal, a correction appended to the answer, or a model-level fix that prevents recurrence? How should providers communicate outcomes to complainants without divulging proprietary details? And how do these processes scale across consumer, enterprise, and open-source deployments?

These are not purely legal questions; they are institutional ones. Providers will differ in size, architecture, and risk tolerance. What is “reasonable” for a small research lab may be impossible for a platform serving hundreds of millions of queries each day. Even a liability regime will have to account for these practical challenges. The goal should be to encourage transparent pathways for correction without freezing product design or privileging incumbents who can afford heavy compliance.

Furthermore, leaving the question on a “case-by-case” isn’t really an ideal scenario, as there could still be lengthy and costly legal proceedings on certain difficult cases. Ideally, the final prescription would avoid a way to adjudicate case by case and instead provide a doctrine which statute can handle without constant judicial review.

There is also the challenge of incentives. A too-easy path to liability invites defensive over-censorship and reduces diversity in model behavior. Conversely, a rule so protective that it forecloses any recourse for fabricated accusations undermines public trust and invites pressure for broad statutory fixes. The plausible middle is a standard that keeps the bar high with specific fabrication, no user inducement, clear notice, and demonstrable harm while making space for targeted remedies when those elements are satisfied.

Technology may even make the question moot in a few years, as hallucinations themselves appear to be declining[4] as providers improve training data, deploy retrieval augmentation, and design inference-time checks that reduce unsupported claims. Tool use and verification steps can now force models to consult authoritative sources rather than guessing. Enterprise deployments increasingly combine models with curated knowledge bases, narrowing the space in which confabulation can occur.

None of this eliminates error, but the direction of travel suggests fewer, not more, pure fabrications over time.

That trend matters for policy design. Building heavy liability regimes around a shrinking problem risks ossifying markets just as engineering improvements are taking hold. A lighter-touch approach that preserves broad speech protections while acknowledging a narrow, targeted remedy for the hardest cases can evolve as the technology does, tightening or relaxing as evidence warrants.

A related dynamic is attribution. As the mix of outputs shifts from free-form generation to synthesis anchored in citations or internal documents, questions of who said what become easier to answer. Where a model accurately repeats a third party's claim, the traditional remedy points to the human author. Where it draws from an employer's corpus, internal governance and contractual remedies can address errors more effectively than public tort law. The narrow lane described above is primarily for the residue: when the model appears to be the original and only source of a defamatory claim, and the user did not ask for or plant it.

Conclusion

The legal system does not need to reinvent free speech principles to accommodate generative AI. A sensible starting point recognizes LLM outputs as speech and treats providers, in the ordinary case, as intermediaries entitled to the same broad protections that have enabled the modern internet. Within that baseline, traditional defamation doctrine can still do work. One plausible route, outlined here as a possibility rather than a prescription, is to reserve a narrow, carefully defined lane for the exceptional case where a model appears to invent a specific defamatory falsehood, the user did not ask for or seed the claim, the provider receives particularized notice, the claimant demonstrates actual harm, and the provider fails to act with reasonable promptness.

That pathway fits the existing architecture without collapsing it. It offers a remedy for real injuries without imposing generalized duties that chill lawful speech and entrench incumbents. It leaves room for markets to reduce error rates through better training, retrieval, verification, and user design. And it acknowledges the outer boundary: non-defamation harms stemming from bad ideas, satire, or user misinterpretation typically remain outside tort's reach and are better addressed through product improvements and consumer choice.

There will be hard cases, and courts will need to calibrate standards with care. But complexity is not an argument for abandoning first principles. The combination of strong speech protection, targeted liability for clearly provable harms, a user-inducement safeguard, and deference to iterative improvement has served the broader internet reasonably well. With cautious adaptation, it can do the same for AI.

Spence Purnell is a senior fellow of technology and innovation at the R Street Institute

ENDNOTES

---

1 “Proposed Amicus Brief in Support of Appeal - Garcia V. Character Technologies, Inc..” The Foundation for Individual Rights and Expression. Accessed October 23, 2025. https://www.thefire.org/research-learn/proposed-amicus-brief-support-appeal-garcia-v-character-technologies-inc

2 Purnell, Spence and Thierer, Adam. “Trump’s ‘Woke AI’ Efforts Should Focus on the Real Problem”. *R Street Institute*. https://www.rstreet.org/commentary/trumps-woke-ai-efforts-should-focus-on-the-real-problem/

3 Cohn, Ari. “Brief of Proposed Amicus Curiae Foundation for Individual Rights and Expression in Support of Character Technologies Motion for Certification of Immediate Appeal”. *Foundation For Individual Rights and Expression*. 06/23/2025.

4 Nielsen, Jakob. “AI Hallucinations on the Decline.” *UX Tigers*. https://www.uxtigers.com/post/ai-hallucinations