The Future Of Gen AI Training Amid Reddit Data Scraping Suit

By **Michael Hobbs and Di'Vennci Lucas** (November 19, 2025)

Michael Hobbs

Di'Vennci Lucas

On Oct. 22, social media platform Reddit sued artificial intelligence startup Perplexity AI, along with three other companies: SerpApi LLC, Oxylabs UAB, and AWMProxy, in the U.S. District Court for the Southern District of New York.

Reddit's lawsuit against Perplexity — Reddit Inc. v. SerpApi LLC — and several scraping/proxy providers is notable for what it is, and what it is not.

Unlike many pending cases against generative AI companies seeking training content that center on copyright infringement and fair use of materials that are otherwise available on the Internet, Reddit's claims focus on how the defendants allegedly obtained Reddit data — the alleged use of false identities, proxies and other antisecurity techniques to scrape at an industrial scale.

So instead of reading like a law school treatise on the future of fair use after the U.S. Supreme Court's 2023 decision in Andy Warhol Foundation for the Visual Arts v. Goldsmith, the Reddit v. Perplexity lawsuit reads like the cyber tactics of a computer super hacker from a movie — well, maybe the most legalese cyber hacker movie ever.

Whether this represents an aberration or the future of generative AI remains to be seen.

In its complaint, Reddit accuses the defendants, collectively, of violating the Digital Millennium Copyright Act's circumvention of technological control measures pursuant to Title 17 of the U.S. Code, Section 1201(a)(1)(A).

While SerpApi and Oxylabs each face additional charges pursuant to the DMCA, SerpApi and Perplexity also face claims of civil conspiracy.

Notably, the complaint frames the dispute as unlawful circumvention and unfair competition, not a classic copyright infringement fight. It also underscores that Reddit's data is already licensed to major AI companies and is available for lawful use — if you pay.

What Makes This Case Different

At bottom, Reddit's complaint is about access, not use.

In Reddit's telling, the defendants allegedly masked identities, disguised web scrapers, hid locations, rotated IPs, forged credentials, ignored robots.txt, and overwhelmed or sidestepped anti-bot defenses — conduct aimed at defeating technical and contractual gatekeeping rather than anything to do with downstream model training or outputs.

Instead of alleging traditional copyright infringement, Reddit anchors its claims in anti-circumvention and trafficking-in-circumvention-technology theories, supplemented by state unfair competition or unjust enrichment laws.

Content licensing is also front and center: Reddit emphasizes that it licenses programmatic access — including to Google and OpenAI — and signals a willingness to license the defendants on commercial terms.

The thrust of the narrative is that the defendants chose to evade those terms and protections with the assertion that they just did not want to take the time or pay the money to get an available license.

The Alleged Conduct

Reddit alleges that the defendants orchestrated industrial scale scraping of Reddit content by pulling Reddit pages from Google's search results rather than accessing Reddit directly, and by defeating both Google's and Reddit's technical defenses.

According to the complaint, the defendants deployed "server-swarms" to mimic human traffic — tactics aimed at evading robots.txt, rate limits, captchas and other anti-bot controls.

Reddit further claims that none of the defendants had authorization or a license to access or use Reddit data in this manner. Instead of using Reddit's data application programming interface under agreed terms, the defendants allegedly bypassed Reddit's licensing program and continued their activity despite notice, including a 2024 cease-and-desist order, with Perplexity allegedly increasing its reliance on Reddit content afterward.

Reddit asserts this conduct undermined its licensing model, damaged user trust and forced significant investment in additional security.

The complaint attributes specific roles:

- SerpApi and Oxylabs are said to market and provide scraping tools and vast proxy networks designed to bypass website protections used by Reddit and others;

- AWMProxy allegedly supplies proxy infrastructure that conceals identity and location for industrial-scale scraping; and

- Perplexity is alleged to have obtained and used the scraped Reddit data — including via SerpApi — for commercial purposes, engaging in stealth behavior to avoid detection.

Reddit's Licensing Posture

Reddit's position is that its data is invaluable to AI companies, especially commercially. It points to existing licensing agreements — including with Google and OpenAI — and a structured data application programming interface program that offers lawful, bulk access under clear terms.

Reddit states that it is willing to license the defendants as well, provided they enter commercial agreements and adhere to guardrails designed to protect users, content integrity and platform reliability.

Reddit further argues that unlicensed scraping undermines its licensing relationships by devaluing paid, compliant access and eroding the incentives for others to honor agreed

protections and fees.

This conduct, in Reddit's view, weakens the sustainability of its licensing model, encourages noncompliance, and forces additional enforcement and security costs while risking user trust and the integrity of the platform.

Why This Matters for AI Companies

A Shift in Legal Exposure

Even if a company avoids, or defends against, infringement or fair use claims, it can still face substantial risk if it acquires training datasets through methods characterized as circumvention, trespass-like conduct or unfair competition.

The path by which data is obtained is legally consequential.

Business Model Pressure

This case spotlights an economic question: Can generative AI businesses sustainably train on licensed datasets at scale, or do they depend on free — and often restricted — content?

The plaintiffs are drawing a bright line: Pay for access, comply with policies, and respect technical controls — or risk injunctions and damages.

Compliance and Provenance Become Differentiators

As licensing pathways expand, investors, enterprise customers, and regulators will look for verifiable data provenance, adherence to robots.txt and site policies, transparent user agents, and auditable ingestion practices.

"Clean" training pipelines may become a competitive advantage and a requirement for partnerships.

Practical Takeaways for AI and Data Ingestion Teams

Audit your acquisition routes.

Map every data source and confirm compliance with site terms, robots.txt and authentication requirements. Avoid indirect scraping via intermediaries that circumvent controls, e.g., SERP scraping at scale.

Use licensed application programming interfaces and contracts.

Where bulk access is needed, pursue commercial licenses and abide by guardrails, i.e., rate limits, use restrictions and privacy protections.

Build technical guardrails.

Enforce robots.txt respect by default, maintain transparent user-agent strings, throttle responsibly, and document consent or authorization. Create a provenance ledger for training sets.

Align product and legal strategy.

If your model depends on high-volume web content, budget for licensing or redesign ingestion to rely on permissible public sources. Be realistic about the cost and timing of licensed datasets.

Conclusion

According to a Wall Street Journal article published May 12, Perplexity raised $500 million, which valued the company at $14 billion.[1]

Yahoo! Finance reported in a story published July 18 that in June, it raised another $100 million, lifting its valuation to $18 billion with the backing of Nvidia, Softbank and others.[2]

At the same time, The Wall Street Journal also reported in an article published Nov. 13 that OpenAI's losses in 2025 could reach $74 billion.[3]

So, with a tremendous upside, but significant expenses and losses, the question remains whether generative AI companies will use their equity to fund licenses with content providers, or their expanding expenses will lead to claims that they are breaking the law to get much-needed content.

Michael D. Hobbs Jr. is a partner and Di'Vennci K. Lucas is an associate at Troutman Pepper Locke LLP.

[1] https://www.wsj.com/tech/ai-startup-perplexitys-valuation-surges-to-14-billion-in-fresh-funding-round.

[2] https://finance.yahoo.com/news/perplexity-ai-achieves-18bn-valuation-113151944.html?.

[3] https://finance.yahoo.com/news/big-tech-soaring-profits-ugly-122500177.html.