

HIGHLIGHTS OF

AI TRAINING DATA

6 CASES TO KNOW



BARTZ V. ANTHROPIC

JURISDICTION

United States

DATES

Fair-use order — 23 Jun 2025

Settlement approved — 20 Jul 2026

PLAINTIFF

Andrea Bartz and a class of authors

DEFENDANT

Anthropic PBC

THE DISPUTE

Authors alleged Anthropic downloaded millions of books from pirate “shadow libraries” such as Library Genesis to build the training corpus behind Claude. Anthropic did not dispute downloading the pirated copies.

PLAINTIFF ARGUED — training on stolen copies infringed regardless of what the training was for.

DEFENDANT ARGUED — acquiring and keeping the copies was “reasonably necessary” to build a training corpus; and training an LLM on the books is, separately, a highly transformative fair use, regardless of how any given copy was sourced.

THE OUTCOME

The court agreed that training itself can be fair use, but held that acquiring and permanently retaining a pirated library, beyond what training required, is a separate, unprotected act. That exposure drove a \$1.5bn class settlement (roughly \$3,000 per work) approved in July 2026, requiring Anthropic to destroy the pirated source files. Objectors also sought deletion or retraining of the AI models built on that data — the court rejected this. Only the pirated library had to go; Claude itself, and the models trained on it, were unaffected.

SOURCES

Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal.), order on summary judgment (23 Jun 2025, Alsup, J.); final approval of class settlement (20 Jul 2026, Martínez-Olguín, J.).

Training an AI model can be fair use. Training on pirated copies cannot.

SONY & WARNER V. ANTHROPIC

JURISDICTION

United States

DATES

Filed 28 Aug 2026

PLAINTIFF

Sony Music Publishing (US) LLC, Warner Chappell Music, Inc., et al.

DEFENDANT

Anthropic PBC

THE DISPUTE

Dozens of Sony and Warner music-publishing companies say Anthropic pirated song lyrics and sheet music from illegal sites to train Claude — and that Claude will still repeat those lyrics word-for-word if you ask it.

PLAINTIFF ARGUED — this is the same kind of piracy Anthropic already got in trouble for in the Bartz case, just with music instead of books. And the filters Anthropic added afterward don't really work — you can still get Claude to output full lyrics by asking a different way.

DEFENDANT ARGUED — no formal answer has been filed yet.

THE OUTCOME

Too early to call. This is the case to watch for two reasons. First, it tests whether the same logic from Bartz — that piracy can sink a fair-use defence even when the training itself is legal — also holds up with a different plaintiff and a different kind of content. Second, it puts a new question in front of the same court: does it matter that the model can still produce protected lyrics today, not just that pirated files were downloaded years ago?

SOURCES

Sony Music Publishing (US) LLC et al. v. Anthropic PBC, No. 5:26-cv-09217 (N.D. Cal., filed 28 Aug 2026).

If your model's output reproduces a copyrighted work, you're in a high-risk zone.

THOMSON REUTERS V. ROSS INTELLIGENCE

JURISDICTION

United States

DATES

District court ruling — 11 Feb 2025

Appeal argued — 11 Jun 2026

Decision pending

PLAINTIFF

Thomson Reuters Enterprise Centre GmbH

DEFENDANT

Ross Intelligence Inc.

THE DISPUTE

Ross trained a competing AI-powered legal search tool on Westlaw's proprietary headnotes, obtained through a licensed intermediary rather than directly from Thomson Reuters.

PLAINTIFF ARGUED — the use was commercial and non-transformative — headnotes fed straight into a rival product — and it threatened Thomson Reuters' own market for licensing content to AI developers.

DEFENDANT ARGUED — building a different kind of search tool from the headnotes was transformative fair use. On market harm specifically, Ross argued no real market for licensing headnotes to train AI existed at the time, so Thomson Reuters could not claim to have lost revenue from a market that didn't yet exist.

THE OUTCOME

The district court rejected the fair-use defence outright because Ross's product competed directly in Westlaw's own core market. On appeal, the dispute has narrowed to a more specific question: beyond that direct competition, does Thomson Reuters also suffer "market harm" from losing a hypothetical licensing market — being paid by AI companies specifically for the right to train on its headnotes — even though no such market really existed before AI came along? The Third Circuit's answer, expected by the end of 2026, will be the first federal appellate ruling on AI-training fair use.

SOURCES

Thomson Reuters Enter. Ctr. GmbH v. Ross Intelligence Inc., No. 1:20-cv-613-SB, 2025 WL 458520 (D. Del. 11 Feb 2025); on appeal, No. 25-2153 (3d Cir., argued 11 Jun 2026).

Training on data to build a rival product is the riskiest move in AI copyright law.

KADREY V. META

JURISDICTION

United States

DATES

Summary judgment order — 25 Jun 2025

PLAINTIFF

Richard Kadrey, Sarah Silverman and other authors

DEFENDANT

Meta Platforms, Inc.

THE DISPUTE

Authors alleged Meta downloaded their books from shadow libraries such as LibGen and Anna’s Archive, without permission, to train the Llama models.

PLAINTIFF ARGUED — training on their books displaced the market for the works and enabled a competing product built on stolen text.

DEFENDANT ARGUED — training is transformative, Llama cannot reproduce meaningful portions of any single work, and no market harm was ever actually shown.

THE OUTCOME

The court granted summary judgment for Meta — but on narrow grounds: the authors simply failed to put evidence of market harm before the court, not because such harm can never exist. The judge went out of his way to endorse “market dilution” — a market flooded with AI-generated substitutes — as a theory that could win a future case if properly evidenced.

SOURCES

Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417-VC (N.D. Cal. 25 Jun 2025) (Chhabria, J.).

This fair-use win turned on a gap in the evidence. It shouldn't be taken as a general rule, though.

GETTY IMAGES V. STABILITY AI

JURISDICTION

United Kingdom

DATES

Judgment handed down — 4 Nov 2025

PLAINTIFF

Getty Images (US), Inc. and related Getty/iStock entities

DEFENDANT

Stability AI Limited

THE DISPUTE

Getty alleged Stable Diffusion was trained on millions of its images without a licence, that the trained model was itself an infringing copy imported into the UK, and that some outputs reproduced Getty/iStock watermarks.

PLAINTIFF ARGUED — the model’s weights embody its copyrighted images, so bringing the model into the UK was secondary infringement, and watermarked outputs infringed its trade marks.

DEFENDANT ARGUED — training happened entirely outside the UK; model weights are statistical parameters, not stored copies of any photograph; and watermark artefacts were rare, historic, and immaterial.

THE OUTCOME

Getty dropped its primary UK training claim once it was clear training occurred abroad, outside the court’s reach. The court rejected secondary infringement outright: model weights are not an “infringing copy” under UK copyright law. The trade mark claim succeeded only on a narrow, historic set of watermark outputs, with no meaningful remedy. Stability AI prevailed on the claims that mattered.

SOURCES

Getty Images (US), Inc. v Stability AI Ltd [2025] EWHC 2863 (Ch) (4 Nov 2025, Joanna Smith J.).

A trained model isn't a copy of its training images. Where you train, and what it outputs, still matter.

NEW YORK TIMES V. OPENAI & MICROSOFT

JURISDICTION

United States

DATES

Filed 27 Dec 2023

Motion to dismiss denied — 26 Mar 2025

DOJ statement of interest — 2 Sep 2026

PLAINTIFF

The New York Times Company (consolidated with other publishers)

DEFENDANT

OpenAI, Inc. and Microsoft Corporation

THE DISPUTE

The Times alleges its articles were used without permission to train GPT models, and that ChatGPT and Copilot can reproduce or closely paraphrase its journalism — including behind its paywall.

PLAINTIFF ARGUED — outputs that substitute for reading the original article cause direct market harm, separate from and in addition to any harm from the training act itself.

DEFENDANT ARGUED — training on published articles is transformative fair use, and the handful of near-verbatim outputs were the product of adversarial prompting, not ordinary use.

THE OUTCOME

The court let the case proceed past the motion to dismiss, allowing discovery on whether output-level substitution can defeat a fair-use defence that the training act alone might have won. On 2 September 2026 the U.S. Department of Justice filed a statement of interest backing OpenAI and Microsoft, calling LLM training "extraordinarily transformative." The DOJ is effectively arguing not just that training is fair use, but that requiring licences would hand a "moat" to today's largest AI companies and archive-rich publishers while shutting out smaller startups, and that higher U.S. training costs would favour overseas AI rivals. The Times called it siding with "a handful of trillion-dollar AI companies"; the DOJ's brief did not address the output-reproduction question the case is actually being litigated over.

SOURCES

The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y.), order on motions to dismiss (26 Mar 2025, Stein, J.); U.S. Statement of Interest (2 Sep 2026).

What comes out of the model can undo a fair-use win on how it was trained.

WHERE THIS LEAVES YOU

Six cases in, a clear framework emerges: courts treat training itself, how the data was acquired, and what the model goes on to output as three separate questions — not one. Four factors keep deciding the outcome:

- **Was the data lawfully acquired?** Courts forgive a lot here — but not hoarding pirated copies indefinitely, beyond what training actually needed.
- **Is the use transformative?** Still AI developers' strongest card.
- **Is there market harm?** Still rightsholders' strongest card.
- **Does the output reproduce the work?** The question now at the centre of the fight.

In 2026 data provenance is a product-strategy question — as much a part of your build as your model choice or your go-to-market.

OUR RECOMMENDED APPROACH



BYTES & RIGHTS —

Bytes & Rights is a team of highly skilled lawyers with deep expertise in digital business and technology.

Drawing from backgrounds at Oxford, top law firms, and Big-4 consulting, we provide legal solutions to fuel your success.

CONTACTS



Tatiana Nikiforova

Partner Bytes&Rights, Magister Juris (Oxford)

+44 79 3424 2123

tatiana.nikiforova@bytes-and-rights.com

Bytes And Rights Ltd is not authorised or regulated by the Solicitors Regulation Authority ("SRA") and does not provide services reserved to authorised persons under the Legal Services Act 2007. While we operate to the same standards of confidentiality, diligence, and ethical conduct as regulated firms, SRA protections (including the Compensation Fund) do not apply, and we do not currently carry professional indemnity insurance.