Federal Judge Rules AI Training on Lawfully Acquired Books Is Fair Use—But Anthropic’s 7 Million Pirated Copies Are Not

A federal judge just declared that AI companies can legally train on books they own—but Anthropic built its library from 7 million pirated titles, and that changes everything.

The Ruling That Splits AI Training Into Legal and Illegal Territories

On June 23, 2025, Judge William Alsup of the U.S. District Court for the Northern District of California issued the first major federal fair use ruling on AI training in Bartz et al. v. Anthropic PBC (Case No. 3:24-cv-05417-WHA). The decision creates a bright legal line that every AI company, startup founder, and CTO building with large language models must now understand.

The core holding is deceptively simple: training LLMs on lawfully acquired books constitutes fair use. Judge Alsup described this training process as “exceedingly transformative,” analogizing it to centuries of human reading and learning. Authors, the court held, “cannot exclude others from using their works to learn.”

But here’s where Anthropic’s legal victory collapses into a $1.5 billion liability: the company built what the court called its “central library” from approximately 7 million copyrighted books downloaded from LibGen and PiLiMi—shadow library piracy sites that exist specifically to circumvent copyright protection. For these works, the court rejected fair use entirely.

The plaintiffs include authors Andrea Bartz (We Were Never Here), Charles Graeber (The Good Nurse), and Kirk Wallace Johnson (The Feather Thief). They represent what the court certified as a class of approximately 500,000 titles with proper ISBN/ASIN identifiers and U.S. Copyright Office registration. A trial on piracy liability was scheduled for December 1, 2025, though the parties have since moved toward a proposed $1.5 billion settlement framework.

The ruling also addressed a subsidiary question that affects data pipeline engineering: digitizing lawfully purchased print books into a single internal digital copy—with the print copy destroyed afterward—constitutes fair use under a format-shifting rationale. This green-lights a specific data acquisition workflow, but only under tightly controlled conditions.

Why This Decision Reshapes the AI Industry’s Legal Calculus

This ruling matters because it simultaneously validates and invalidates core practices across the AI industry. The decision is not a blanket permission slip for AI training, nor is it a death sentence for the technology. It’s a surgical distinction that will force companies to audit their data provenance with unprecedented rigor.

Winners: Companies With Clean Data Pipelines

Organizations that invested in legitimate data acquisition—licensed datasets, books purchased at scale, partnerships with publishers, content created in-house—now have federal judicial backing for their approach. The “exceedingly transformative” language is powerful precedent. It means the output of training (an LLM’s learned

capabilities) is so different from the input (the books themselves) that copying for training purposes doesn’t compete with the original works’ market.

This validates what companies like OpenAI have argued theoretically but never had confirmed judicially: that machine learning on copyrighted works is fundamentally different from copying those works for distribution. The court explicitly compared LLM training to human learning, which copyright law has never restricted. You can read a book, learn from it, and write something new without paying royalties to every author who influenced your thinking.

Losers: Anyone Who Took Shortcuts on Data Acquisition

The 7-million-book piracy finding exposes a dirty secret of the AI industry: many companies built their training corpora from sources they knew (or should have known) were illegally compiled. LibGen and PiLiMi are not obscure corners of the internet. They are well-known piracy operations that have faced repeated legal action. Any engineer or executive who approved ingesting data from these sources took a calculated risk that just materialized into a potential nine-figure liability.

The certified class of 500,000 registered titles represents only works with proper documentation. The total exposure could be larger if the court later expands the class definition or if individual authors bring separate claims.

The Middle Ground: Companies Who Don’t Know Their Data Origins

Perhaps the most uncomfortable position belongs to organizations that genuinely cannot trace where their training data originated. If your data pipeline ingested text from aggregators, scraped websites, or third-party datasets without verifying provenance, you cannot confidently assert you avoided pirated sources. This ruling transforms data lineage from a nice-to-have compliance checkbox into a litigation defense requirement.

Technical Implications for AI System Architecture

The Bartz ruling forces technical decisions that go beyond legal compliance. The architecture of your data pipeline, the documentation of your training process, and the auditability of your model development all become potential evidence in future

litigation.

Data Provenance as a First-Class Engineering Concern

Building training datasets now requires the same chain-of-custody documentation that regulated industries demand for other sensitive processes. Every text file in your training corpus should have:

- **Source verification:** Where did this content originate? Can you prove it?
- **Acquisition documentation:** How was it obtained? Purchase receipt? License agreement? API terms of service?
- **Rights confirmation:** What rights does the source actually have to distribute this content?
- **Timestamped ingestion logs:** When did this enter your pipeline, and who approved it?

The format-shifting holding also creates a specific compliant workflow: purchase physical books, digitize them to a single internal copy, destroy the physical originals. This is operationally expensive but legally defensible. Companies building training infrastructure should consider whether dedicated book-acquisition and digitization pipelines make economic sense at scale.

The LibGen Problem Is Bigger Than Anthropic

LibGen appears in numerous public datasets that researchers and companies have used for training. The Debevoise analysis of the Anthropic and Meta decisions notes that contamination from pirated sources may be more widespread than many organizations realize.

If your organization used any of the following without careful filtering, you may have LibGen-derived content in your training data:

- Common Crawl (which has indexed LibGen mirrors)
- Books3 or similar book-focused datasets
- Academic paper corpora that weren't vetted for source legitimacy
- Any "comprehensive" text dataset compiled by third parties

The technical challenge is that once pirated content enters a training corpus, it's difficult to identify and impossible to "untrain" without retraining from scratch.

Model archaeology—determining what data influenced a model’s training—remains an unsolved research problem. This means companies with potentially contaminated models face a choice between expensive retraining or uncertain legal exposure.

Differential Treatment of Data Sources

The ruling suggests that AI companies should implement tiered data handling based on provenance confidence:

Tier 1 (High Confidence): Content you created, licensed with explicit AI training rights, or purchased directly from publishers with documented chain of custody. Use freely for training.

Tier 2 (Medium Confidence): Content obtained from platforms with terms of service that permit derivative use, or public domain works with verified status. Use with documentation, prepare to defend.

Tier 3 (Low Confidence): Content from aggregators, scraped sources, or datasets you didn’t compile yourself. Audit thoroughly before use; consider exclusion.

Tier 4 (Prohibited): Content from known piracy sources, circumvention of access controls, or terms-of-service violations. Never use.

What the Headlines Are Getting Wrong

Media coverage of Bartz v. Anthropic has largely framed this as a split decision—some win for AI companies, some win for authors. This framing misses the structural implications of what Judge Alsup actually held.

The “Learning” Analogy Is Stronger Than It Appears

Most commentary has treated the comparison between LLM training and human learning as a rhetorical flourish. It’s not. It’s a legal holding with precedential weight. Judge Alsup didn’t say training is “like” learning as a metaphor; he said the legal principle that protects human learning from copyright claims applies equally to machine learning.

This is a significant expansion of fair use doctrine. Previous fair use cases in the

technology context—Google Books, HathiTrust, the various search engine cases—involved transformative use of copies for indexing or reference purposes. The outputs were search results or snippets that pointed users back to original works.

Bartz goes further. It holds that the transformation occurs in the training process itself, not in the output. The LLM learns statistical patterns from text, but it does not reproduce that text—and this learning is protected regardless of whether users ever access the original works. This collapses the distinction between intermediate copying (which courts have sometimes permitted) and final-product copying (which they scrutinize more heavily).

The Piracy Distinction Isn't Really About Piracy

The court's rejection of fair use for LibGen-sourced books isn't actually about piracy per se. It's about the first fair use factor: the purpose and character of the use.

When Anthropic obtained books through purchase, its purpose was research and development of a new technology. When it downloaded books from LibGen, its purpose was to avoid paying for content it knew was copyrighted. The second purpose taints the entire use, regardless of what Anthropic did with the books afterward.

This distinction matters for companies considering gray-area data sources. The question isn't just "is this content copyrighted?" The question is "why are we obtaining it this way instead of the legitimate way?" If the answer is "because it's free" or "because the legitimate way is too expensive," you've just failed the first fair use factor before you even analyze the other three.

The Tension With Kadrey v. Meta Is Overblown

Coverage of the Bartz decision has emphasized the apparent conflict with Judge Chhabria's ruling in Kadrey v. Meta, issued just days later, which found Meta's Llama training to be fair use even though some pirated sources were involved.

But the cases are distinguishable on their facts. In Kadrey, the pirated works were a smaller fraction of the total training corpus, and Meta had stronger arguments about the transformative nature of its specific use. The courts aren't necessarily in conflict; they're applying the same multi-factor test to different factual records.

The real lesson is that fair use remains a fact-intensive inquiry. There's no safe harbor for "we only pirated a little bit." But there's also no automatic liability for any contact with potentially problematic sources. The proportion, the company's knowledge, the availability of legitimate alternatives, and the commercial nature of the use all matter.

What CTOs and Technical Leaders Should Do Now

The Bartz ruling isn't a theoretical exercise for law review articles. It demands concrete responses from anyone building or operating AI systems trained on text.

Audit Your Training Data Immediately

If you have not already conducted a comprehensive audit of your training data sources, start now. You need to answer these questions with documentation:

- What datasets did you use for training or fine-tuning?
- For each dataset, who compiled it and under what terms?
- Did any dataset include content from known piracy sources?
- Can you identify and quantify potentially problematic content?
- Do you have records of when you acquired each dataset and from whom?

This audit should involve both engineering and legal teams. Engineers understand what data went where; lawyers understand what questions will be asked in discovery.

Implement Forward-Looking Data Governance

For new training runs and model updates, implement a formal data governance process:

Approval workflows: No dataset enters your training pipeline without documented approval from someone authorized to accept legal risk.

Source verification: Require evidence of legitimate origin for all text data. "We downloaded it from the internet" is not sufficient documentation.

Exclusion lists: Maintain an updated list of sources you will not ingest, including known piracy sites, scraped platforms with prohibitive ToS, and datasets with

documented provenance problems.

Retention policies: Keep records of what went into each model version. If you need to demonstrate that a specific model was trained only on clean data, you need contemporaneous documentation.

Consider Data Licensing Investments

The Bartz ruling strengthens the negotiating position of content licensors. Publishers, authors' organizations, and content aggregators with clear rights can now point to federal precedent establishing that clean provenance matters.

This creates both a cost and an opportunity. Companies that invest in legitimate data licensing—whether through direct publisher relationships, platforms like the Wikimedia Enterprise API, or specialized data vendors—can compete on compliance as a differentiator. Companies that continue to rely on questionable sources face escalating legal risk and potential exclusion from risk-averse enterprise customers.

The $1.5 billion settlement framework in Bartz also suggests that the cost of getting caught far exceeds the cost of doing it right initially. Seven million books at $214 per book in implicit damages is a painful lesson in false economy.

Evaluate Your Model Lineage

If you're using foundation models from third parties—whether open-source releases or API-based services—you inherit their data provenance risks. The Bartz and Kadrey decisions apply to the organizations that trained the models, but downstream users may face secondary liability theories or contractual exposure.

For critical applications, consider:

- What training data disclosures has your model provider made?
- Do your licensing agreements include representations about training data legality?
- What indemnification provisions protect you if the base model faces litigation?
- Should you require data provenance attestations from vendors?

The AFS Law analysis notes that these decisions will influence how enterprise buyers evaluate AI vendors. Compliance-sensitive industries—financial services,

healthcare, government contracting—may begin requiring training data documentation as a vendor qualification criterion.

Where AI Copyright Law Heads From Here

The Bartz ruling is not the final word on AI training and copyright. It’s a district court decision that will almost certainly be appealed, and it addresses only a subset of the pending AI copyright cases. But it establishes a framework that will shape the next 12 to 18 months of legal development.

The Settlement as Precedent

The proposed $1.5 billion settlement, if finalized and approved, creates an informal valuation framework for future AI copyright disputes. At roughly $3,000 per title for 500,000 registered works, it suggests that copyright holders can extract meaningful compensation for unauthorized training use—but not the billions-per-company figures that some plaintiffs have claimed.

This matters because it gives both sides a reference point for negotiations. Authors and publishers know what the market will bear. AI companies know what their exposure looks like. This should accelerate settlements in other pending cases, including the various actions against OpenAI, Microsoft, and Stability AI.

Legislative and Regulatory Responses

Congress has shown increasing interest in AI copyright issues, and the Bartz ruling will inform legislative debate. The decision validates the existing fair use framework as capable of handling AI-specific questions, which reduces pressure for AI-specific copyright legislation. But the piracy holding also demonstrates enforcement gaps that legislators may seek to address.

Expect to see proposals for:

- Mandatory training data disclosure requirements
- Enhanced penalties for AI-specific copyright infringement
- Safe harbor provisions for companies that meet certain data governance standards
- Compulsory licensing schemes for AI training (similar to music licensing models)

The EU’s AI Act already includes training data documentation requirements, and the Bartz ruling may encourage similar provisions in U.S. legislation or regulation.

The Next Frontier: Synthetic Data and Data Attribution

As the legal risks of training on copyrighted content become clearer, expect increased investment in alternatives. Synthetic data generation—using AI to create training data rather than collecting it from human-created sources—becomes more attractive despite its technical limitations.

Similarly, research into data attribution and influence measurement becomes legally valuable, not just scientifically interesting. If you can demonstrate that specific content had minimal influence on your model’s outputs, you may have stronger fair use arguments.

Companies developing these technical capabilities—provable data provenance, influence quantification, clean-room training methodologies—will find a market among compliance-conscious organizations seeking to minimize litigation risk.

International Implications

The Bartz ruling applies only in the United States, but it influences global AI development because most leading AI companies operate under U.S. jurisdiction. Non-U.S. companies training models on copyrighted content and serving U.S. customers may face exposure even if their home jurisdictions don’t recognize similar claims.

The ruling also creates potential regulatory arbitrage. Jurisdictions with weaker copyright enforcement may become attractive for AI training operations, though this creates its own risks as models must eventually enter commerce somewhere.

The Tweetable Takeaways

For those who need the executive summary:

On fair use: “Training on books you own is legal learning. Judge Alsup just compared LLM training to centuries of human reading—and said both are protected.”

On piracy: “Seven million pirated books, $1.5 billion settlement. The false economy of free data just got a price tag.”

On compliance: “Data provenance moved from ‘engineering hygiene’ to ‘litigation evidence’ in one ruling. Audit your sources before plaintiffs do.”

On the future: “This isn’t the end of AI copyright disputes. It’s the beginning of a framework for resolving them.”

The Strategic Imperative

Bartz v. Anthropic clarifies that the AI industry’s legal exposure stems not from training on copyrighted works—which is transformative and legal—but from how companies acquired those works. The distinction is between learning and theft, and courts will enforce it.

For technical leaders, this means data governance is no longer optional. Every company building or deploying AI systems needs documented, defensible data acquisition practices. The cost of compliance is real but finite. The cost of getting caught, as Anthropic is discovering, is measured in billions.

The companies that will thrive in the post-Bartz environment are those that treat data provenance as a competitive advantage rather than a compliance burden. Clean data pipelines enable faster model development, lower legal risk, and stronger positioning with enterprise customers who cannot afford to inherit their vendors’ copyright problems.

The Bartz ruling doesn’t change what AI can do—it clarifies what AI companies must do to build it legally, and the price they’ll pay if they don’t.