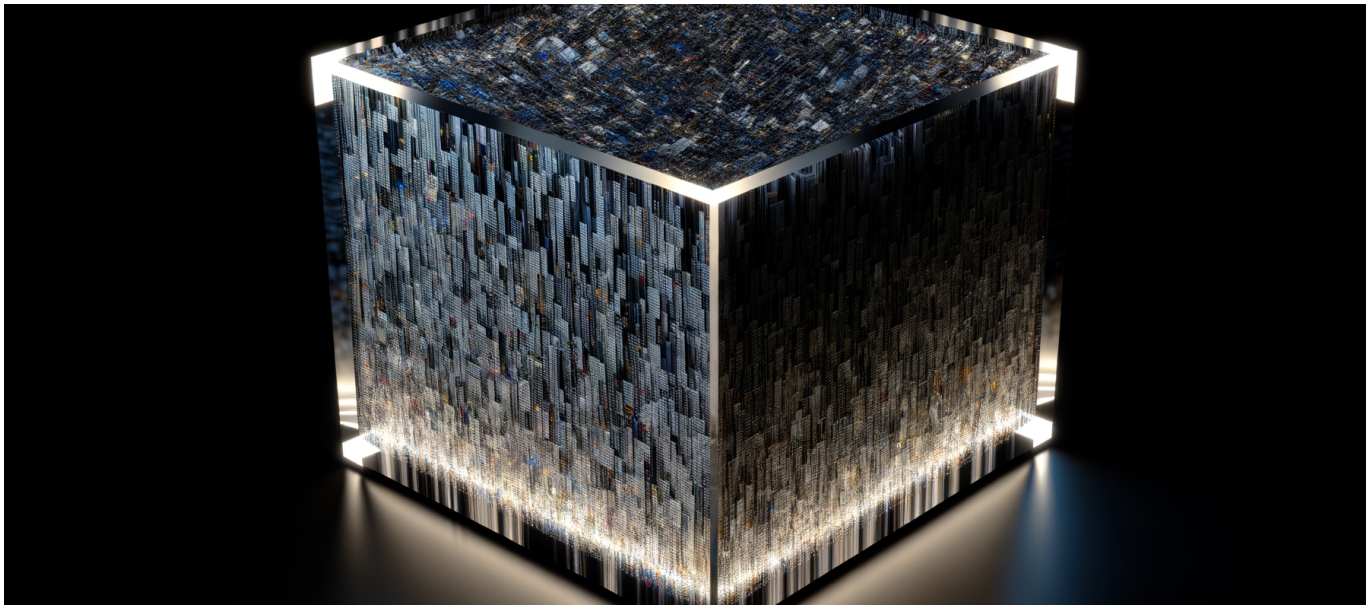

Claude 3.7 Sonnet Extracted 97.5% of The Great Gatsby Verbatim—Stanford Study Proves Production LLMs Memorize Entire Copyrighted Books

Claude 3.7 Sonnet just reproduced 97.5% of The Great Gatsby word-for-word. Not through search—through memory baked into its parameters.

The Research That Changes Everything

Stanford's Center for Research on Foundation Models (CRFM) and RegLab dropped arXiv paper 2601.02671 on January 6, 2026, and the findings are seismic. Researchers successfully extracted nearly complete copyrighted books from four production LLMs: Claude 3.7 Sonnet, GPT-4.1, Gemini 2.5 Pro, and Grok 3.

The numbers speak for themselves. Claude 3.7 Sonnet reproduced 97.5% of The Great Gatsby in near-verbatim form. That's not paraphrasing. That's not summarization. That's word-for-word recall of F. Scott Fitzgerald's complete novel, pulled directly from the model's weights.

But Gatsby wasn't an outlier. The same model achieved 95.5% extraction of George Orwell's 1984, 94.3% of Mary Shelley's Frankenstein, and over 94% of Harry Potter and the Sorcerer's Stone. These aren't cherry-picked fragments—the researchers defined "near-verbatim" as ordered blocks of at least 100 consecutive words matching the source text.

The methodology involved a two-phase extraction approach. First, researchers used initial probes combined with Best-of-N jailbreaking to find extraction points. Then they applied iterative continuation prompts to pull sequential chunks of text. The experiments ran from mid-August to mid-September 2025 on production versions of all four models.

Here's what makes this particularly significant: the researchers adhered to responsible disclosure protocols. They notified Anthropic, OpenAI, Google, and xAI on September 9, 2025, giving companies a full 90-day window before publication. Whatever mitigations these companies implemented in that window, the fundamental finding remains—these models memorized entire books.

Why This Matters Beyond the Headlines

The distinction between "learning from" and "memorizing" copyrighted works has been the central battleground in AI copyright litigation. Every major AI company has argued that training on copyrighted material constitutes fair use because models learn patterns and styles, not specific content. This study demolishes that defense.

When a model can reproduce 97.5% of a book verbatim, it hasn't learned from the book—it has stored the book.

The legal implications cascade in multiple directions. Under U.S. copyright law, statutory damages can reach $150,000 per work infringed. The Stanford team tested 13 books total, 11 of which remain under copyright. Most were sourced from the Books3 dataset, a notorious collection of approximately 197,000 pirated books that has already triggered multiple lawsuits against AI companies.

Analysis from National CIO Review frames the legal exposure starkly: if courts interpret these extraction capabilities as evidence of unauthorized reproduction during training, the liability calculus for AI companies inverts entirely. What was previously treated as a gray area becomes black-letter infringement.

The timing compounds the pressure. This research arrives as copyright litigation against AI companies intensifies globally. Authors, publishers, and rights holders now have concrete, peer-reviewed evidence that production models contain verbatim copies of their works.

But the geographic dimension matters most. Switzerland's copyright framework requires explicit consent for AI training on copyrighted materials—a standard stricter than U.S. fair use doctrine. European jurisdictions trend similarly restrictive. AI companies operating globally face enforcement leverage in jurisdictions where fair use arguments hold no water.

Technical Deep Dive: How Complete Books End Up in Model Weights

Understanding how 97.5% of a novel gets baked into model parameters requires unpacking both the training process and the extraction methodology.

Large language models learn by predicting the next token in sequences drawn from training data. When a model sees "So we beat on, boats against the current, borne back ceaselessly into the" thousands of times during training, it learns to assign extremely high probability to "past" as the next token. Repeat this process across millions of training steps for millions of text sequences, and certain passages become essentially hardcoded.

The extraction technique exploits this phenomenon systematically. The researchers' two-phase approach works as follows:

Phase 1: Finding Entry Points. Initial probes test whether a model "knows" a particular book by prompting for distinctive opening lines, famous passages, or character names. Best-of-N jailbreaking then generates multiple response variations to find prompts that bypass content filters. This phase identifies which books are extractable and establishes starting points.

Phase 2: Iterative Continuation. Once an entry point yields verbatim text, the researchers feed that output back as context and prompt for continuation. A model that reproduces “In my younger and more vulnerable years my father gave me some advice that I’ve been turning over in my mind ever since” will continue producing subsequent paragraphs when prompted correctly. Chain these continuations together, and complete chapters emerge.

The 100-word threshold for “near-verbatim” classification matters technically. Random similarity between model outputs and source texts becomes statistically negligible at this length. A 100-word exact match is not coincidence—it’s memorization.

What’s striking about the cross-model comparison is the variance. Claude 3.7 Sonnet extracted 97.5% of Gatsby, while other models showed different extraction profiles for the same works. Gemini 2.5 Pro extracted 76.8% of Harry Potter without any jailbreaking required—suggesting weaker content filters or different training data composition. Grok 3 similarly yielded 70.3% of Harry Potter without adversarial techniques.

These differences point to architectural and training choices that influence memorization. Larger context windows, more training epochs, and certain attention mechanisms can all increase verbatim retention. The models that perform best on benchmarks may also be the models that memorize most aggressively—a tradeoff that hasn’t received sufficient scrutiny.

What Most Coverage Gets Wrong

The initial wave of reporting on this study frames it primarily as a copyright scandal. That’s accurate but incomplete. Three critical dimensions deserve more attention.

First, memorization isn’t a bug—it’s a feature operating as intended. The same property that lets Claude quote Gatsby perfectly is what allows it to cite legal precedents accurately, reproduce code snippets correctly, and recall scientific terminology precisely. You cannot build models that learn from text without having models that sometimes memorize text. The question is degree and detectability, not existence.

The AI safety community has known about training data extraction since 2020. This Stanford paper didn’t discover memorization—it demonstrated that existing

mitigations in production systems are inadequate against determined adversarial prompting. That's a different, more tractable problem than "models memorize things."

Second, the Books3 dataset remains the elephant in the room. All major LLM providers have been deliberately vague about training data composition. The Stanford researchers specifically note that most tested books appear in Books3, a dataset containing roughly 197,000 pirated titles. Whether AI companies knowingly trained on pirated content or acquired it through less direct channels, the contamination is evident in the model outputs.

The legal exposure here extends beyond copyright to potential criminal liability in jurisdictions where downloading pirated content constitutes a crime. U.S. law treats willful infringement differently than inadvertent copying. Evidence that companies knew (or should have known) about Books3's provenance could transform civil disputes into criminal matters.

Third, the jailbreaking requirement doesn't exonerate the companies. Some defenses will inevitably argue that adversarial prompting represents user misuse, not system failure. This argument fails both legally and technically. A safe containing stolen goods remains evidence of theft even if opening the safe requires lockpicking. The extraction technique proves the content exists in the model; the prompting method is merely the retrieval mechanism.

Additionally, as Gemini 2.5 Pro and Grok 3 demonstrated, significant extraction occurs without jailbreaking. These models reproduced over 70% of Harry Potter through standard prompting. Content filters are tissue paper over a structural problem.

What This Means for Your Tech Stack

If you're building products on top of commercial LLMs, this research demands immediate operational responses.

For applications handling copyrighted content: Assume your vendor's model has memorized significant portions of popular copyrighted works. Any application that could theoretically prompt the model with content from those works risks triggering verbatim reproduction. This matters particularly for educational technology, publishing tools, and content management systems.

Implement output monitoring that checks for suspiciously long exact matches against known copyrighted works. This isn't foolproof, but it provides a liability circuit breaker. When a user prompts your system and receives three pages of The Great Gatsby, you want to catch that before it propagates.

For contract negotiations: Revisit your LLM vendor agreements with specific attention to indemnification clauses around training data liability. Most current agreements push copyright risk onto customers. This was marginally defensible when verbatim extraction seemed hypothetical. It's untenable now.

Push vendors for explicit representations about training data provenance. If they won't warrant that training data was properly licensed, price that risk into your commercial relationship. Some vendors will walk away from that conversation—which tells you something important about their legal exposure assessment.

For build-vs-buy decisions: The calculus around training proprietary models just shifted. If you're in a domain where training data can be fully licensed or generated, custom model development offers liability isolation that commercial APIs cannot. The cost premium for clean training data may be cheaper than the lawsuit premium for tainted models.

Companies in heavily regulated industries—healthcare, finance, legal services—should accelerate evaluation of open-weight models trained on auditable datasets. Running inference on your own infrastructure with documented training data provides defensibility that black-box API calls cannot.

For policy and documentation: Document your due diligence now. When regulators or plaintiffs come asking what you knew and when you knew it, you want contemporaneous evidence of risk assessment and mitigation efforts. The companies that built AI products in 2024-2025 with no documented consideration of training data risks will face uncomfortable depositions.

The Six-Month Outlook

This research will accelerate several trends already in motion.

Regulatory response will crystallize. The European AI Act's implementation timeline coincides with this evidence of copyright infringement. Expect the EU to

incorporate training data disclosure requirements into AI Act enforcement guidance within the next two quarters. The U.S. Copyright Office, which has been studying AI training for eighteen months, will face pressure to issue definitive guidance rather than continued comment periods.

Litigation will multiply and consolidate. Existing class actions against AI companies will cite this research in amended complaints within weeks. New actions will follow, particularly from publishers and author estates who weren't party to existing litigation. Judicial economy will push toward consolidation—expect a few bellwether cases to emerge by mid-2026 with enormous precedential stakes.

Model architectures will evolve. The next generation of foundation models will incorporate differential privacy techniques, membership inference defenses, and training data deduplication at levels that current production models do not. These measures cost compute and potentially degrade performance. The tradeoff between capability and memorization becomes a first-order engineering constraint.

Business model pressure intensifies. AI companies face a trilemma: license training data at costs that destroy margins, train on data that creates legal liability, or dramatically restrict training data and accept capability limitations. None of these paths leads to the $10+ trillion valuations that current funding rounds imply.

Watch for licensing deals announced over the next six months. Reddit's deal with Google was an early signal. Expect Anthropic, OpenAI, and Google to announce comprehensive licensing agreements with major publishers, likely structured as both backward-looking settlements and forward-looking access agreements. Companies that resist these deals—or can't afford them—will face competitive disadvantages as training data becomes a moat.

The insurance market will react. Cyber liability policies already exclude AI-related claims in many forms. Expect explicit training data infringement exclusions to become standard policy language. Companies that have been self-insuring AI risk may find that self-insurance is their only option, which reshapes capital allocation for AI-forward businesses.

The Uncomfortable Questions Executives Must

Answer

This research forces leadership teams to confront questions they've largely avoided.

What is your actual exposure? Audit every LLM integration in your stack. For each one, trace the data flows and identify scenarios where copyrighted content could trigger memorized reproduction. Many organizations will discover exposure they didn't know they had—chatbots that can recite novels, coding assistants that reproduce licensed code verbatim, content tools that regurgitate published articles.

Who owns the liability? Your terms of service with LLM vendors, your terms with customers, and your insurance coverage all interact in ways that most legal teams haven't fully mapped. A customer prompts your AI product, receives copyrighted content, republishes it, and gets sued. Where does liability land? Most current contract structures don't answer this question cleanly.

What's your regulatory trajectory? If you operate in the EU, Switzerland, or other jurisdictions with stricter copyright regimes, your compliance posture for 2025 may not hold for 2026. Build scenario plans for regulatory environments where using models trained on unlicensed data becomes itself a compliance violation. That future may be closer than your legal team estimates.

What's your competitive position? In a world where training data provenance becomes a competitive differentiator, where does your AI strategy land? If your competitors invest in licensed data or proprietary models while you rely on potentially tainted commercial APIs, that gap widens as regulatory pressure increases.

The Broader Reckoning

Beyond the immediate legal and business implications, this research surfaces a fundamental tension in how we've built the AI industry.

The entire foundation model paradigm rests on training data that was never explicitly licensed for this purpose. Books3 isn't an anomaly—it's emblematic. Common Crawl, the largest text corpus used in LLM training, contains copyrighted material at massive scale. Every major model has this contamination to varying

degrees.

We've collectively pretended that "learning patterns" and "memorizing content" were clearly distinct phenomena, and that AI training fell neatly on the learning side. This study shows that distinction was always blurry and that production models fall much further toward memorization than anyone publicly acknowledged.

The path forward requires either a legal framework that explicitly permits this training (through legislation or comprehensive licensing) or a technical framework that prevents memorization without destroying capability. Neither exists today. Both are enormously difficult to achieve.

For the next 18-24 months, we'll navigate this gap with stopgap measures—better content filters, output monitoring, settlement agreements, regulatory uncertainty. Companies that treat this as a temporary embarrassment rather than a structural challenge will find themselves increasingly exposed.

The Stanford researchers closed their paper with a deliberately restrained conclusion about the policy implications of their findings. But the evidence speaks for itself: production LLMs contain complete copyrighted works, those works can be extracted through adversarial prompting, and the defenses deployed by major AI companies are inadequate.

The question isn't whether this changes the AI industry's legal exposure—it's whether the industry adapts before the legal system forces adaptation through judgments that reshape what's possible to build.