

Chapter 04: AI Arrives When the Human Raw Material Already Exists

From Surveillance Capitalism to AI Imperialism | Chapter 04: AI arrives when the human raw material already exists

0.1. Abstract

Foundation models do not emerge from nowhere. The U.S. Once captured human production can be recombined into proprietary weights, the surveillance economy acquires a new layer.

0.2. Keywords

AI arrives when the human raw material already exists; surveillance capitalism; AI imperialism; AI weights; patrimony of humanity; evidence-based dossier

0.3. Research Question

How does ai arrives when the human raw material already exists function within the dossier's larger argument, and what does the documentary record allow this chapter to establish?

0.4. Scope and Framing Note

This paper examines ai arrives when the human raw material already exists within the broader dossier *From Surveillance Capitalism to AI Imperialism*. It is intentionally written to circulate on its own: necessary context is restated, evidence boundaries are made explicit, and cited material is reproduced in paper-local form rather than delegated to repo navigation. It contributes to the series as a paper that provides the central evidentiary basis for treating frontier weights as a patrimony and governance problem rather than a neutral market outcome. Adjacent reinforcement appears in Chapter 00, Chapter 05, Chapter 06, Chapter 10, but those cross-references are supportive rather than required for basic comprehension.

0.5. Central Thesis

Foundation models do not emerge from nowhere. They emerge after a long period in which human expression, cultural production, and socially filtered judgment were captured at scale, weakly

compensated if at all, and converted into machine-trainable inputs.¹

0.6. Evidence and Method Note

This paper is derived from the chapter corpus for Chapter 04 and is grounded in the project's structured research OS. It relies on source notes, extracted claims, entity profiles, and timeline events already normalized under the dossier's evidence hierarchy. Documented facts are asserted directly where the record supports them; disputed matters are marked as such; interpretive claims are bounded explicitly; and speculative overreach is isolated in the evidence-boundary appendix.

0.7. Introduction

This paper asks: How does ai arrives when the human raw material already exists function within the dossier's larger argument, and what does the documentary record allow this chapter to establish? Its working answer is clear from the start: Foundation models do not emerge from nowhere. They emerge after a long period in which human expression, cultural production, and socially filtered judgment were captured at scale, weakly compensated if at all, and converted into machine-trainable inputs.² The objective is not only to recount developments, but to show why the subject of this chapter belongs inside a structural argument about extraction, enclosure, legitimacy, and reform.

0.8. Historical or Institutional Context

The U.S. Copyright Office has already treated AI training on copyrighted works as a major live policy problem rather than a settled technical inevitability. The 2023 inquiry and 2025 report together show that training-input legality, disclosure, and compensation were contested near the start of the generative-AI boom and remain contested now.³

0.9. Main Analysis

The technical and legal record also undermines the industry's preferred abstraction. Research on memorization shows that large language models can reproduce identifiable material from training data, while court records and legal scholarship now separate at least two layers of conflict: training-use doctrine and input-acquisition legality.⁴

The social-to-model bridge is also explicit. GPT-2's WebText was built from socially filtered outbound links shared on Reddit, and by February 2024 Reddit had moved from diffuse public conver-

¹Sources: U.S. Copyright Office, Copyright and Artificial Intelligence... (2025-05-09; see Appendix A1); OpenAI, Language Models are Unsupervised Mult... (2019-02-14; see Appendix A2); AP, Reddit strikes \$60M deal allowing Goo... (2024-02-22; see Appendix A3)

²Sources: U.S. Copyright Office, Copyright and Artificial Intelligence... (2025-05-09; see Appendix A1); OpenAI, Language Models are Unsupervised Mult... (2019-02-14; see Appendix A2); AP, Reddit strikes \$60M deal allowing Goo... (2024-02-22; see Appendix A3)

³Sources: GovInfo, Copyright and Artificial Intelligence... (2023-08-30; see Appendix A4); U.S. Copyright Office, Copyright and Artificial Intelligence... (2025-05-09; see Appendix A1)

⁴Sources: arXiv, Extracting Training Data from Large L... (2020-12-14; see Appendix A5); arXiv, Generative AI Training and Copyright Law (2025-02-21; see Appendix A6); U.S. District Court for the Northern District of California via CourtListener, Bartz et al. v. Anthropic PBC, Order... (2025-06-23; see Appendix A7); U.S. District Court for the Southern District of New York via Justia, The New York Times Company v. Microso... (2025-04-04; see Appendix A8)

sation to direct licensing, striking a major agreement that allowed Google to use Reddit posts for AI training and related product improvement. The extraction economy did not end with unpaid capture; it evolved into direct licensing over accumulated human expression.⁵

The labor layer matters too. Research on annotation and outsourced human labor shows that model building depends not only on scraped creative works but also on significant human verification, correction, and operational support that is often geographically externalized and weakly valued.⁶

Taken together, the record shows a sequence in which collectively produced expression becomes training input, archived conversation becomes licensable model fuel, and additional human labor is used to clean, annotate, and operationalize the resulting systems. Value that first appeared as diffuse public production is progressively reorganized inside proprietary model pipelines.⁷

Interpretation: the strategic move is not only model creation but control of the interface through which machine-packaged intelligence is accessed. In that reading, the shelf is the interface: the prompt box, chatbot, or subscription layer becomes the first point of contact between a buyer and capability distilled from human work. Once that first contact is captured, value can be repriced downward and dependence can deepen even where total technical replacement remains incomplete.⁸

The House AI Task Force report gives this chapter an institutional bridge beyond agencies and courts. By December 17, 2024, a bipartisan congressional report was already recording the training conflict in language overlapping with the dossier’s core concerns about commercialization, consent, credit, compensation, and derivative-work theory.⁹

Interpretation: this is the dossier’s primary evidentiary base for **AI Weights as Patrimony of Humanity**. If frontier weights are built from collectively produced human expression and labor, then they cannot be treated as ordinary private assets by default. The strongest normative direction pushes toward broad, open, utility-like availability under public-interest rules rather than permanent private scarcity. That is advocated doctrine, not settled law.¹⁰

0.10. Position Within the Series

This paper sits inside a coordinated 11-paper architecture. Its nearest companions are Chapter 00 (Opening framework), Chapter 05 (Talent, overhiring, valuation, and layoffs), Chapter 06 (AI as a dual-use rhetorical weapon), Chapter 10 (What to do). In that series logic, Chapter 04 provides

⁵Sources: OpenAI, Language Models are Unsupervised Mult... (2019-02-14; see Appendix A2); AP, Reddit strikes \$60M deal allowing Goo... (2024-02-22; see Appendix A3)

⁶Sources: Big Data & Society, The problem with annotation. Human la... (2023-07-25; see Appendix A9)

⁷Sources: OpenAI, Language Models are Unsupervised Mult... (2019-02-14; see Appendix A2); AP, Reddit strikes \$60M deal allowing Goo... (2024-02-22; see Appendix A3); Big Data & Society, The problem with annotation. Human la... (2023-07-25; see Appendix A9)

⁸Sources: OpenAI, Language Models are Unsupervised Mult... (2019-02-14; see Appendix A2); AP, Reddit strikes \$60M deal allowing Goo... (2024-02-22; see Appendix A3); Big Data & Society, The problem with annotation. Human la... (2023-07-25; see Appendix A9)

⁹Sources: U.S. House of Representatives, Bipartisan House Task Force Report on... (2024-12-17; see Appendix A10); TwinChat, AI Copyright Weights: A New Frontier... (n.d.; see Appendix A11)

¹⁰Sources: TwinChat, AI Copyright Weights: A New Frontier... (n.d.; see Appendix A11); U.S. Copyright Office, Copyright and Artificial Intelligence... (2025-05-09; see Appendix A1); U.S. House of Representatives, Bipartisan House Task Force Report on... (2024-12-17; see Appendix A10)

the central evidentiary basis for treating frontier weights as a patrimony and governance problem rather than a neutral market outcome.

0.11. Counterarguments

The current record is strong on conflict, appropriation structure, memorization, licensing, and the social-to-model bridge. It is still weaker on full training-stack provenance across every frontier lab. This chapter should therefore challenge enclosure forcefully without pretending that every corpus is already fully mapped.¹¹

0.12. Limits of the Evidence

The current record is strong on conflict, appropriation structure, memorization, licensing, and the social-to-model bridge. It is still weaker on full training-stack provenance across every frontier lab. This chapter should therefore challenge enclosure forcefully without pretending that every corpus is already fully mapped.¹²

0.13. Reform Relevance

Once captured human production can be recombined into proprietary weights, the surveillance economy acquires a new layer. It no longer only predicts behavior. It begins to substitute for the producers whose accumulated traces made the system possible in the first place. That is why the ownership question has to expand into a governance question.¹³

0.14. Conclusion

Once captured human production can be recombined into proprietary weights, the surveillance economy acquires a new layer. The next question is how that enclosure story meets labor, finance, and layoffs. In the architecture of this series, Chapter 04 therefore functions as a self-contained argument while also advancing the cumulative path toward the dossier's three reforms.

0.15. Notes

This paper is designed to circulate independently. Public notes point readers to the real external documents first, then to the paper-local appendix entry that explains why each source matters.

The underlying research OS distinguishes among **Documented Fact**, **Disputed Fact**, **Hypothesis / Interpretation**, and **Speculative Narrative Risk**. In Chapter 04, those boundaries remain visible in the prose, in the bibliography, and in the appendix sections that summarize claims and caution points.

¹¹Sources: U.S. Copyright Office, Copyright and Artificial Intelligence... (2025-05-09; see Appendix A1); U.S. District Court for the Southern District of New York via Justia, The New York Times Company v. Microso... (2025-04-04; see Appendix A8)

¹²Sources: U.S. Copyright Office, Copyright and Artificial Intelligence... (2025-05-09; see Appendix A1); U.S. District Court for the Southern District of New York via Justia, The New York Times Company v. Microso... (2025-04-04; see Appendix A8)

¹³Sources: OpenAI, Language Models are Unsupervised Mult... (2019-02-14; see Appendix A2); AP, Reddit strikes \$60M deal allowing Goo... (2024-02-22; see Appendix A3); TwinChat, AI Copyright Weights: A New Frontier... (n.d.; see Appendix A11)

0.16. References / Bibliography

1. U.S. Copyright Office. Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-publication Version). 2025-05-09. T1 report. Paper appendix: Appendix A1. Corpus companion entry: Source S-126. Audit ID: `usco-ai-training-report-2025`. Relevance: That the U.S. Copyright Office treats AI training on copyrighted works as a major live legal and policy problem rather than a settled matter.
2. OpenAI. Language Models are Unsupervised Multitask Learners. 2019-02-14. T1 paper. Paper appendix: Appendix A2. Corpus companion entry: Source S-108. Audit ID: `radford-gpt2-webtext-2019`. Relevance: That a major OpenAI training corpus was explicitly built from socially filtered outbound links shared on Reddit.
3. AP. Reddit strikes \$60M deal allowing Google to train AI models on its posts, unveils IPO plans. 2024-02-22. T2 news_article. Paper appendix: Appendix A3. Corpus companion entry: Source S-024. Audit ID: `ap-reddit-google-ai-training-deal-2024`. Relevance: That by February 2024 a major social platform was explicitly licensing accumulated user conversation data for frontier-AI training.
4. GovInfo. Copyright and Artificial Intelligence; Notice of Inquiry. 2023-08-30. T1 federal_register_notice. Paper appendix: Appendix A4. Corpus companion entry: Source S-125. Audit ID: `usco-ai-noi-2023`. Relevance: That by August 2023 the U.S. Copyright Office had already elevated AI training inputs and disclosure questions into a formal federal inquiry.
5. arXiv. Extracting Training Data from Large Language Models. 2020-12-14. T1 paper. Paper appendix: Appendix A5. Corpus companion entry: Source S-038. Audit ID: `carlini-extracting-training-data-2020`. Relevance: That large language models can sometimes reproduce identifiable verbatim material from their training data rather than only abstract statistical patterns.
6. arXiv. Generative AI Training and Copyright Law. 2025-02-21. T1 paper. Paper appendix: Appendix A6. Corpus companion entry: Source S-051. Audit ID: `dornis-generative-ai-training-copyright`. Relevance: That there is serious academic legal scholarship directly challenging the idea that generative AI training automatically falls under familiar exceptions.
7. U.S. District Court for the Northern District of California via CourtListener. Bartz et al. v. Anthropic PBC, Order on Fair Use. 2025-06-23. T1 court_order. Paper appendix: Appendix A7. Corpus companion entry: Source S-035. Audit ID: `bartz-anthropic-fair-use-order-2025`. Relevance: That the strongest primary legal record already separates the question of model training from the question of how the training corpus was assembled.
8. U.S. District Court for the Southern District of New York via Justia. The New York Times Company v. Microsoft Corporation et al, Opinion on Motions to Dismiss. 2025-04-04. T1 court_order. Paper appendix: Appendix A8. Corpus companion entry: Source S-085. Audit ID: `nyt-openai-motion-dismiss-order-2025`. Relevance: That another federal court outside the Anthropic case treated training-stage copying and dataset assembly as central issues serious enough to survive dismissal.
9. Big Data & Society. The problem with annotation. Human labour and outsourcing between France and Madagascar. 2023-07-25. T1 paper. Paper appendix: Appendix A9. Corpus companion entry: Source S-077. Audit ID: `le-ludec-cornet-casilli-annotation-madagascar-2023`. Relevance: That AI systems depend on significant human labor for data generation, annotation, verification, and operational correction.

10. U.S. House of Representatives. Bipartisan House Task Force Report on Artificial Intelligence. 2024-12-17. T1 report. Paper appendix: Appendix A10. Corpus companion entry: Source S-070. Audit ID: `house-ai-task-force-report-2024`. Relevance: That the dossier's copyright / weights / compensation framing entered a bipartisan U.S. congressional policy record by December 17, 2024.
11. TwinChat. AI Copyright Weights: A New Frontier in Intellectual Property Law. n.d.. T3 essay. Paper appendix: Appendix A11. Corpus companion entry: Source S-122. Audit ID: `twinchat-ai-copyright-weights`. Relevance: That the dossier has a coherent thesis about how weights condense previously captured human labor.

0.17. Appendix A. Source Register for This Paper

0.17.1. Appendix A1. Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-publication Version)

Source citation: Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-publication Version) **Institution / author:** U.S. Copyright Office **Date:** 2025-05-09 **Tier / type:** T1 / report **Why it matters in this paper:** That the U.S. Copyright Office treats AI training on copyrighted works as a major live legal and policy problem rather than a settled matter. **Corpus companion entry:** Source S-126 **Audit ID:** `usco-ai-training-report-2025`

0.17.2. Appendix A2. Language Models are Unsupervised Multitask Learners

Source citation: Language Models are Unsupervised Multitask Learners **Institution / author:** OpenAI **Date:** 2019-02-14 **Tier / type:** T1 / paper **Why it matters in this paper:** That a major OpenAI training corpus was explicitly built from socially filtered outbound links shared on Reddit. **Corpus companion entry:** Source S-108 **Audit ID:** `radford-gpt2-webtext-2019`

0.17.3. Appendix A3. Reddit strikes \$60M deal allowing Google to train AI models on its posts, unveils IPO plans

Source citation: Reddit strikes \$60M deal allowing Google to train AI models on its posts, unveils IPO plans **Institution / author:** AP **Date:** 2024-02-22 **Tier / type:** T2 / news_article **Why it matters in this paper:** That by February 2024 a major social platform was explicitly licensing accumulated user conversation data for frontier-AI training. **Corpus companion entry:** Source S-024 **Audit ID:** `ap-reddit-google-ai-training-deal-2024`

0.17.4. Appendix A4. Copyright and Artificial Intelligence; Notice of Inquiry

Source citation: Copyright and Artificial Intelligence; Notice of Inquiry **Institution / author:** GovInfo **Date:** 2023-08-30 **Tier / type:** T1 / federal_register_notice **Why it matters in this paper:** That by August 2023 the U.S. Copyright Office had already elevated AI training inputs and disclosure questions into a formal federal inquiry. **Corpus companion entry:** Source S-125 **Audit ID:** `usco-ai-noi-2023`

0.17.5. Appendix A5. Extracting Training Data from Large Language Models

Source citation: Extracting Training Data from Large Language Models **Institution / author:** arXiv **Date:** 2020-12-14 **Tier / type:** T1 / paper **Why it matters in this paper:** That large language models can sometimes reproduce identifiable verbatim material from their training data rather than only abstract statistical patterns. **Corpus companion entry:** Source S-038 **Audit ID:** carlini-extracting-training-data-2020

0.17.6. Appendix A6. Generative AI Training and Copyright Law

Source citation: Generative AI Training and Copyright Law **Institution / author:** arXiv **Date:** 2025-02-21 **Tier / type:** T1 / paper **Why it matters in this paper:** That there is serious academic legal scholarship directly challenging the idea that generative AI training automatically falls under familiar exceptions. **Corpus companion entry:** Source S-051 **Audit ID:** dornis-generative-ai-training-copyright-2025

0.17.7. Appendix A7. Bartz et al. v. Anthropic PBC, Order on Fair Use

Source citation: Bartz et al. v. Anthropic PBC, Order on Fair Use **Institution / author:** U.S. District Court for the Northern District of California via CourtListener **Date:** 2025-06-23 **Tier / type:** T1 / court_order **Why it matters in this paper:** That the strongest primary legal record already separates the question of model training from the question of how the training corpus was assembled. **Corpus companion entry:** Source S-035 **Audit ID:** bartz-anthropic-fair-use-order-2025

0.17.8. Appendix A8. The New York Times Company v. Microsoft Corporation et al, Opinion on Motions to Dismiss

Source citation: The New York Times Company v. Microsoft Corporation et al, Opinion on Motions to Dismiss **Institution / author:** U.S. District Court for the Southern District of New York via Justia **Date:** 2025-04-04 **Tier / type:** T1 / court_order **Why it matters in this paper:** That another federal court outside the Anthropic case treated training-stage copying and dataset assembly as central issues serious enough to survive dismissal. **Corpus companion entry:** Source S-085 **Audit ID:** nyt-openai-motion-dismiss-order-2025

0.17.9. Appendix A9. The problem with annotation. Human labour and outsourcing between France and Madagascar

Source citation: The problem with annotation. Human labour and outsourcing between France and Madagascar **Institution / author:** Big Data & Society **Date:** 2023-07-25 **Tier / type:** T1 / paper **Why it matters in this paper:** That AI systems depend on significant human labor for data generation, annotation, verification, and operational correction. **Corpus companion entry:** Source S-077 **Audit ID:** le-ludec-cornet-casilli-annotation-madagascar-2023

0.17.10. Appendix A10. Bipartisan House Task Force Report on Artificial Intelligence

Source citation: Bipartisan House Task Force Report on Artificial Intelligence **Institution / author:** U.S. House of Representatives **Date:** 2024-12-17 **Tier / type:** T1 / report **Why it matters in this paper:** That the dossier's copyright / weights / compensation framing entered a bipartisan U.S. congressional policy record by December 17, 2024. **Corpus companion entry:** Source S-070 **Audit ID:** house-ai-task-force-report-2024

0.17.11. Appendix A11. AI Copyright Weights: A New Frontier in Intellectual Property Law

Source citation: AI Copyright Weights: A New Frontier in Intellectual Property Law **Institution / author:** TwinChat **Date:** n.d. **Tier / type:** T3 / essay **Why it matters in this paper:** That the dossier has a coherent thesis about how weights condense previously captured human labor. **Corpus companion entry:** Source S-122 **Audit ID:** twinchat-ai-copyright-weights

0.18. Appendix B. Claims Used in This Paper**0.18.1. Appendix B1. By February 2024 a major social platform was explicitly licensing accumulated use...**

Claim statement: That by February 2024 a major social platform was explicitly licensing accumulated user conversation data for frontier-AI training. **Evidence status:** Documented Fact **Source support:** Appendix A3 **Corpus companion entry:** Claim C-073 **Audit Claim ID:** ap-reddit-google-ai-training-deal-2024-04-data-to-models-copyright-weights-claim-01

0.18.2. Appendix B2. Chapter 4 can extend the social-data-to-models bridge beyond GPT-2-era WebText in...

Claim statement: That Chapter 4 can extend the social-data-to-models bridge beyond GPT-2-era WebText into the later commercial frontier-model period. **Evidence status:** Documented Fact **Source support:** Appendix A3 **Corpus companion entry:** Claim C-074 **Audit Claim ID:** ap-reddit-google-ai-training-deal-2024-04-data-to-models-copyright-weights-claim-02

0.18.3. Appendix B3. The extraction economy did not stop at unpaid capture; it evolved into direct lic...

Claim statement: That the extraction economy did not stop at unpaid capture; it evolved into direct licensing arrangements over archived human expression. **Evidence status:** Documented Fact **Source support:** Appendix A3 **Corpus companion entry:** Claim C-075 **Audit Claim ID:** ap-reddit-google-ai-training-deal-2024-04-data-to-models-copyright-weights-claim-03

0.18.4. Appendix B4. The strongest primary legal record already separates the question of model traini...

Claim statement: That the strongest primary legal record already separates the question of model training from the question of how the training corpus was assembled. **Evidence status:** Docu-

mented Fact **Source support:** Appendix A7 **Corpus companion entry:** Claim C-104 **Audit Claim ID:** bartz-anthropic-fair-use-order-2025-04-data-to-models-copyright-weights-claim-01

0.18.5. Appendix B5. By June 23, 2025, a federal court had accepted a fair-use defense for Anthropic’s...

Claim statement: That by June 23, 2025, a federal court had accepted a fair-use defense for Anthropic’s training on lawfully acquired books while preserving exposure for piracy-related conduct.

Evidence status: Documented Fact **Source support:** Appendix A7 **Corpus companion entry:** Claim C-105 **Audit Claim ID:** bartz-anthropic-fair-use-order-2025-04-data-to-models-copyright-weights-claim-02

0.18.6. Appendix B6. Chapter 4 can rely on a primary court order, not only press summaries, when argui...

Claim statement: That Chapter 4 can rely on a primary court order, not only press summaries, when arguing that AI copyright conflict has split into doctrinal and supply-chain layers.

Evidence status: Documented Fact **Source support:** Appendix A7 **Corpus companion entry:** Claim C-106 **Audit Claim ID:** bartz-anthropic-fair-use-order-2025-04-data-to-models-copyright-weights-claim-03

0.18.7. Appendix B7. Large language models can sometimes reproduce identifiable verbatim material from...

Claim statement: That large language models can sometimes reproduce identifiable verbatim material from their training data rather than only abstract statistical patterns.

Evidence status: Documented Fact **Source support:** Appendix A5 **Corpus companion entry:** Claim C-114 **Audit Claim ID:** carlini-extracting-training-data-2020-04-data-to-models-copyright-weights-claim-04

0.18.8. Appendix B8. Chapter 4 can connect training-data appropriation to memorization and leakage wit...

Claim statement: That Chapter 4 can connect training-data appropriation to memorization and leakage with direct technical evidence.

Evidence status: Documented Fact **Source support:** Appendix A5 **Corpus companion entry:** Claim C-115 **Audit Claim ID:** carlini-extracting-training-data-2020-04-data-to-models-copyright-weights-claim-05

0.18.9. Appendix B9. The notion of model weights as a harmlessly abstract transformation of source mat...

Claim statement: That the notion of model weights as a harmlessly abstract transformation of source material is too simple for the full technical record.

Evidence status: Documented Fact **Source support:** Appendix A5 **Corpus companion entry:** Claim C-116 **Audit Claim ID:** carlini-extracting-training-data-2020-04-data-to-models-copyright-weights-claim-06

0.18.10. Appendix B10. There is serious academic legal scholarship directly challenging the idea that ge...

Claim statement: That there is serious academic legal scholarship directly challenging the idea that generative AI training automatically falls under familiar exceptions. **Evidence status:** Documented Fact

Source support: Appendix A6 **Corpus companion entry:** Claim C-153 **Audit**

Claim ID: dornis-generative-ai-training-copyright-2025-04-data-to-models-copyright-weights-clai

0.18.11. Appendix B11. Memorization is not only a technical curiosity but also part of the copyright debate

Claim statement: That memorization is not only a technical curiosity but also part of the copyright debate. **Evidence status:** Documented Fact **Source support:** Appendix A6 **Corpus com-**

panion entry: Claim C-154 **Audit Claim ID:** dornis-generative-ai-training-copyright-2025-04-data-to

0.18.12. Appendix B12. Chapter 4 can present the legal status of training as contested and actively theo...

Claim statement: That Chapter 4 can present the legal status of training as contested and actively theorized rather than settled by industry analogy. **Evidence status:** Documented Fact

Source support: Appendix A6 **Corpus companion entry:** Claim C-155 **Audit Claim ID:**

dornis-generative-ai-training-copyright-2025-04-data-to-models-copyright-weights-claim-03

0.18.13. Appendix B13. The dossier's copyright / weights / compensation framing entered a bipartisan U.S...

Claim statement: That the dossier's copyright / weights / compensation framing entered a bipartisan U.S. congressional policy record by December 17, 2024. **Evidence status:** Documented Fact

Source support: Appendix A10 **Corpus companion entry:** Claim C-219 **Audit Claim**

ID: house-ai-task-force-report-2024-04-data-to-models-copyright-weights-claim-01

0.18.14. Appendix B14. A formal House report recorded the AI-training dispute in terms that directly ove...

Claim statement: That a formal House report recorded the AI-training dispute in terms that directly overlap with the dossier's concern about reproduction, derivative-work theory, commercialization, consent, credit, and compensation. **Evidence status:** Documented Fact

Source support: Appendix A10 **Corpus companion entry:** Claim C-220 **Audit Claim ID:**

house-ai-task-force-report-2024-04-data-to-models-copyright-weights-claim-02

0.18.15. Appendix B15. Chapter 00 and Chapter 04 can cite institutional uptake of the argument without p...

Claim statement: That Chapter 00 and Chapter 04 can cite institutional uptake of the argument without pretending that congressional citation settles the underlying legal merits. **Evidence status:** Documented Fact

Source support: Appendix A10 **Corpus companion entry:** Claim C-

221 **Audit Claim ID:** house-ai-task-force-report-2024-04-data-to-models-copyright-weights-claim-03

0.18.16. Appendix B16. AI systems depend on significant human labor for data generation, annotation, ver...

Claim statement: That AI systems depend on significant human labor for data generation, annotation, verification, and operational correction. **Evidence status:** Documented Fact

Source support: Appendix A9 **Corpus companion entry:** Claim C-246 **Audit Claim ID:**

le-ludec-cornet-casilli-annotation-madagascar-2023-04-data-to-models-copyright-weights-claim-0

0.18.17. Appendix B17. Chapter 4 can connect model building to undervalued and geographically externaliz...

Claim statement: That Chapter 4 can connect model building to undervalued and geographically externalized human work, not only to scraped copyrighted material. **Evidence status:** Documented Fact

Source support: Appendix A9 **Corpus companion entry:** Claim C-247 **Audit**

Claim ID: le-ludec-cornet-casilli-annotation-madagascar-2023-04-data-to-models-copyright-weight

0.18.18. Appendix B18. A serious research literature already frames AI production as a labor-reorganizat...

Claim statement: That a serious research literature already frames AI production as a labor-reorganization process rather than a purely autonomous technical breakthrough. **Evidence status:** Documented Fact

Source support: Appendix A9 **Corpus companion entry:** Claim C-248 **Au-**

dit Claim ID: le-ludec-cornet-casilli-annotation-madagascar-2023-04-data-to-models-copyright-we

0.18.19. Appendix B19. Another federal court outside the Anthropic case treated training-stage copying a...

Claim statement: That another federal court outside the Anthropic case treated training-stage copying and dataset assembly as central issues serious enough to survive dismissal. **Evidence sta-**

tus: Documented Fact **Source support:** Appendix A8 **Corpus companion entry:** Claim C-280

Audit Claim ID: nyt-openai-motion-dismiss-order-2025-04-data-to-models-copyright-weights-claim-

0.18.20. Appendix B20. Chapter 4 can rely on more than one primary court record when arguing that the AI...

Claim statement: That Chapter 4 can rely on more than one primary court record when arguing that the AI copyright conflict has split into doctrinal and supply-chain layers. **Evidence status:** Documented Fact

Source support: Appendix A8 **Corpus companion entry:** Claim C-281 **Au-**

dit Claim ID: nyt-openai-motion-dismiss-order-2025-04-data-to-models-copyright-weights-claim-02

0.18.21. Appendix B21. The legal record does not treat training as a purely abstract statistical process...

Claim statement: That the legal record does not treat training as a purely abstract statistical process detached from source acquisition and storage practices. **Evidence status:** Documented Fact

Source support: Appendix A8 **Corpus companion entry:** Claim C-282 **Audit Claim ID:**

nyt-openai-motion-dismiss-order-2025-04-data-to-models-copyright-weights-claim-03

0.18.22. Appendix B22. A major OpenAI training corpus was explicitly built from socially filtered outbou...

Claim statement: That a major OpenAI training corpus was explicitly built from socially filtered outbound links shared on Reddit. **Evidence status:** Documented Fact **Source support:** Appendix A2 **Corpus companion entry:** Claim C-356 **Audit Claim ID:** radford-gpt2-webtext-2019-04-data-to-models-copyright-weights-claim-01

0.18.23. Appendix B23. Chapter 4 can make a direct bridge from social-platform activity to model-trainin...

Claim statement: That Chapter 4 can make a direct bridge from social-platform activity to model-training inputs instead of speaking only in broad abstractions about “the internet.” **Evidence status:** Documented Fact **Source support:** Appendix A2 **Corpus companion entry:** Claim C-357 **Audit Claim ID:** radford-gpt2-webtext-2019-04-data-to-models-copyright-weights-claim-02

0.18.24. Appendix B24. Training-data pipelines can incorporate unpaid human curation signals even when t...

Claim statement: That training-data pipelines can incorporate unpaid human curation signals even when the final corpus is presented as a neutral technical dataset. **Evidence status:** Documented Fact **Source support:** Appendix A2 **Corpus companion entry:** Claim C-358 **Audit Claim ID:** radford-gpt2-webtext-2019-04-data-to-models-copyright-weights-claim-03

0.18.25. Appendix B25. The dossier has a coherent thesis about how weights condense previously captured...

Claim statement: That the dossier has a coherent thesis about how weights condense previously captured human labor. **Evidence status:** Documented Fact **Source support:** Appendix A11 **Corpus companion entry:** Claim C-397 **Audit Claim ID:** twinchat-ai-copyright-weights-claim-01

0.18.26. Appendix B26. There is a clear argumentative line linking AI, property, and enclosure

Claim statement: That there is a clear argumentative line linking AI, property, and enclosure. **Evidence status:** Documented Fact **Source support:** Appendix A11 **Corpus companion entry:** Claim C-398 **Audit Claim ID:** twinchat-ai-copyright-weights-claim-02

0.18.27. Appendix B27. By August 2023 the U.S. Copyright Office had already elevated AI training inputs...

Claim statement: That by August 2023 the U.S. Copyright Office had already elevated AI training inputs and disclosure questions into a formal federal inquiry. **Evidence status:** Documented Fact **Source support:** Appendix A4 **Corpus companion entry:** Claim C-407 **Audit Claim ID:** usco-ai-noi-2023-04-data-to-models-copyright-weights-claim-01

0.18.28. Appendix B28. The legal and policy debate around training data was not retrofitted later; it wa...

Claim statement: That the legal and policy debate around training data was not retrofitted later; it was a live official issue near the start of the generative AI boom. **Evidence status:** Documented Fact **Source support:** Appendix A4 **Corpus companion entry:** Claim C-408 **Audit Claim ID:** usco-ai-noi-2023-04-data-to-models-copyright-weights-claim-02

0.18.29. Appendix B29. Chapter 4 can document a formal institutional path from open questions about AI t...

Claim statement: That Chapter 4 can document a formal institutional path from open questions about AI training to later official analysis and policy conflict. **Evidence status:** Documented Fact **Source support:** Appendix A4 **Corpus companion entry:** Claim C-409 **Audit Claim ID:** usco-ai-noi-2023-04-data-to-models-copyright-weights-claim-03

0.18.30. Appendix B30. The U.S. Copyright Office treats AI training on copyrighted works as a major live...

Claim statement: That the U.S. Copyright Office treats AI training on copyrighted works as a major live legal and policy problem rather than a settled matter. **Evidence status:** Documented Fact **Source support:** Appendix A1 **Corpus companion entry:** Claim C-410 **Audit Claim ID:** usco-ai-training-report-2025-04-data-to-models-copyright-weights-claim-01

0.18.31. Appendix B31. Official analysis recognizes copyrighted creative works as central inputs to gene...

Claim statement: That official analysis recognizes copyrighted creative works as central inputs to generative AI development. **Evidence status:** Documented Fact **Source support:** Appendix A1 **Corpus companion entry:** Claim C-411 **Audit Claim ID:** usco-ai-training-report-2025-04-data-to-models-copyright-weights-claim-02

0.18.32. Appendix B32. Chapter 4 can ground its argument about training-data appropriation and licensing...

Claim statement: That Chapter 4 can ground its argument about training-data appropriation and licensing conflict in a primary U.S. government source. **Evidence status:** Documented Fact **Source support:** Appendix A1 **Corpus companion entry:** Claim C-412 **Audit Claim ID:** usco-ai-training-report-2025-04-data-to-models-copyright-weights-claim-03

0.18.33. Appendix B33. Whether existing copyright and fair-use doctrine can absorb frontier-model traini...

Claim statement: That whether existing copyright and fair-use doctrine can absorb frontier-model training without consent or compensation remains actively contested across agencies, courts, and litigants. **Evidence status:** Disputed Fact **Source support:** Appendix A1, Appendix A7 **Corpus companion entry:** Claim C-428 **Audit Claim ID:** vector04-training-legality-contested-claim-

0.18.34. Appendix B34. Dominant intermediaries can consolidate value and pricing power by controlling th...

Claim statement: That dominant intermediaries can consolidate value and pricing power by controlling the access layer through which users reach collectively produced goods, even when they did not produce the underlying material. **Evidence status:** Hypothesis / Interpretation **Source support:** Appendix A2, Appendix A3, Appendix A9 **Corpus companion entry:** Claim C-424 **Audit Claim ID:** vector04-access-layer-power-claim-01

0.18.35. Appendix B35. The AI prompt box, chatbot, or subscription layer can be read as the new shelf or...

Claim statement: That the AI prompt box, chatbot, or subscription layer can be read as the new shelf or point of sale of intelligence because it mediates access to capabilities condensed from collective human production. **Evidence status:** Hypothesis / Interpretation **Source support:** Appendix A2, Appendix A3, Appendix A9 **Corpus companion entry:** Claim C-425 **Audit Claim ID:** vector04-interface-as-shelf-claim-01

0.18.36. Appendix B36. Frontier AI weights can be read as condensed claims over collectively produced hu...

Claim statement: That frontier AI weights can be read as condensed claims over collectively produced human expression and therefore justify patrimony-oriented governance. **Evidence status:** Hypothesis / Interpretation **Source support:** Appendix A11, Appendix A1, Appendix A10 **Corpus companion entry:** Claim C-429 **Audit Claim ID:** vector04-weights-patrimony-governance-claim-01

0.18.37. Appendix B37. Current legal uncertainty or congressional uptake already proves frontier AI weig...

Claim statement: That current legal uncertainty or congressional uptake already proves frontier AI weights are legally settled as patrimony, public trust property, or common heritage. **Evidence status:** Speculative Narrative Risk **Source support:** Appendix A10, Appendix A1 **Corpus companion entry:** Claim C-426 **Audit Claim ID:** vector04-patrimony-already-settled-overclaim-claim-01

0.18.38. Appendix B38. The supermarket / shelf analogy is itself established as a documented fact by the...

Claim statement: That the supermarket / shelf analogy is itself established as a documented fact by the legal or regulatory record rather than as a bounded interpretation built on that record. **Evidence status:** Speculative Narrative Risk **Source support:** Appendix A1, Appendix A3, Appendix A9 **Corpus companion entry:** Claim C-427 **Audit Claim ID:** vector04-shelf-analogy-overproof-claim-01

0.19. Appendix C. Timeline Slice

0.19.1. Appendix C1. 2019-02-14: OpenAI documents a Reddit-filtered corpus for GPT-2 training

Event summary: OpenAI documents a Reddit-filtered corpus for GPT-2 training **Event date:** 2019-02-14 **Source support:** Appendix A2 **Corpus companion entry:** Event T-005 **Audit Event ID:** event-gpt2-webtext-reddit-2019-02-14

0.19.2. Appendix C2. 2023-07-25: A peer-reviewed study maps outsourced annotation labor in AI production chains

Event summary: A peer-reviewed study maps outsourced annotation labor in AI production chains **Event date:** 2023-07-25 **Source support:** Appendix A9 **Corpus companion entry:** Event T-027 **Audit Event ID:** event-annotation-madagascar-paper-2023-07-25

0.19.3. Appendix C3. 2023-08-30: The U.S. Copyright Office opens a formal inquiry into AI training and copyright

Event summary: The U.S. Copyright Office opens a formal inquiry into AI training and copyright **Event date:** 2023-08-30 **Source support:** Appendix A4 **Corpus companion entry:** Event T-028 **Audit Event ID:** event-usco-ai-noi-2023-08-30

0.19.4. Appendix C4. 2024-02-22: Reddit announces a \$60 million agreement allowing Google to train AI models on Reddit...

Event summary: Reddit announces a \$60 million agreement allowing Google to train AI models on Reddit posts **Event date:** 2024-02-22 **Source support:** Appendix A3 **Corpus companion entry:** Event T-043 **Audit Event ID:** event-reddit-google-training-deal-2024-02-22

0.19.5. Appendix C5. 2024-12-17: The Bipartisan House AI Task Force report cites Hassan Uriostegui's weights-and-copyr...

Event summary: The Bipartisan House AI Task Force report cites Hassan Uriostegui's weights-and-copyright argument **Event date:** 2024-12-17 **Source support:** Appendix A10 **Corpus companion entry:** Event T-062 **Audit Event ID:** event-house-ai-task-force-report-2024-12-17

0.19.6. Appendix C6. 2025-04-04: A federal court lets core New York Times copyright claims against OpenAI and Microsof...

Event summary: A federal court lets core New York Times copyright claims against OpenAI and Microsoft proceed **Event date:** 2025-04-04 **Source support:** Appendix A8 **Corpus companion entry:** Event T-073 **Audit Event ID:** event-nyt-openai-dismissal-order-2025-04-04

0.19.7. Appendix C7. 2025-05-09: The U.S. Copyright Office releases its pre-publication report on generative AI training

Event summary: The U.S. Copyright Office releases its pre-publication report on generative AI training **Event date:** 2025-05-09 **Source support:** Appendix A1 **Corpus companion entry:** Event T-075 **Audit Event ID:** event-usco-ai-training-report-2025-05-09

0.19.8. Appendix C8. 2025-06-23: Judge Alsup’s order separates AI training fair use from pirate-library exposure

Event summary: Judge Alsup’s order separates AI training fair use from pirate-library exposure **Event date:** 2025-06-23 **Source support:** Appendix A7 **Corpus companion entry:** Event T-080 **Audit Event ID:** event-bartz-anthropic-order-2025-06-23

0.20. Appendix D. Relevant Entities

0.20.1. Appendix D1. Anthropic

Entity type: company **Role in this paper:** Actor relevant to 04_data_to_models_copyright_weights, 06_ai_double_use_reasoning_deepseek, 07_export_controls_compute_defense_access, 08_power_networks_legitimacy_capture and others. **Relevant sources in this paper:** Appendix A7 **Corpus companion entry:** Entity E-008 **Audit Entity ID:** org-anthropic

0.20.2. Appendix D2. Google

Entity type: company **Role in this paper:** Actor relevant to 04_data_to_models_copyright_weights. **Relevant sources in this paper:** Appendix A3 **Corpus companion entry:** Entity E-018 **Audit Entity ID:** org-google

0.20.3. Appendix D3. OpenAI

Entity type: company **Role in this paper:** Actor relevant to 04_data_to_models_copyright_weights, 06_ai_double_use_reasoning_deepseek, 07_export_controls_compute_defense_access, 08_power_networks_legitimacy_capture and others. **Relevant sources in this paper:** Appendix A8, Appendix A2 **Corpus companion entry:** Entity E-027 **Audit Entity ID:** org-openai

0.20.4. Appendix D4. U.S. Copyright Office

Entity type: government_agency **Role in this paper:** Actor relevant to 04_data_to_models_copyright_weights. **Relevant sources in this paper:** Appendix A4, Appendix A1 **Corpus companion entry:** Entity E-037 **Audit Entity ID:** org-us-copyright-office

0.20.5. Appendix D5. Hassan Uriostegui

Entity type: person **Role in this paper:** Actor relevant to 04_data_to_models_copyright_weights, 08_power_networks_legitimacy_capture. **Relevant sources in this paper:** Appendix A11, Appendix A10 **Corpus companion entry:** Entity E-048 **Audit Entity ID:** person-hassan-uriostegui

0.20.6. Appendix D6. Reddit

Entity type: platform **Role in this paper:** Actor relevant to 02_social_decay_youth_harm_hypersexualization 04_data_to_models_copyright_weights. **Relevant sources in this paper:** Appendix A3, Appendix A2 **Corpus companion entry:** Entity E-032 **Audit Entity ID:** org-reddit

0.21. Appendix E. Evidence Boundaries

0.21.1. Appendix E1. Disputed Matters

- Appendix B33: That whether existing copyright and fair-use doctrine can absorb frontier-model training without consent or compensation remains actively contested across agencies, courts, and litigants.

0.21.2. Appendix E2. Interpretive Boundaries

- Appendix B34: That dominant intermediaries can consolidate value and pricing power by controlling the access layer through which users reach collectively produced goods, even when they did not produce the underlying material.
- Appendix B35: That the AI prompt box, chatbot, or subscription layer can be read as the new shelf or point of sale of intelligence because it mediates access to capabilities condensed from collective human production.
- Appendix B36: That frontier AI weights can be read as condensed claims over collectively produced human expression and therefore justify patrimony-oriented governance.

0.21.3. Appendix E3. Speculative Narrative Risks

- Appendix B37: That current legal uncertainty or congressional uptake already proves frontier AI weights are legally settled as patrimony, public trust property, or common heritage.
- Appendix B38: That the supermarket / shelf analogy is itself established as a documented fact by the legal or regulatory record rather than as a bounded interpretation built on that record.

0.21.4. Appendix E4. Sensitive or Review-Worthy Note Flags

- Appendix A3. Companion note: Note N-027. Risk flag: review. Note focus: That by February 2024 a major social platform was explicitly licensing accumulated user conversation data for frontier-AI training.
- Appendix A7. Companion note: Note N-028. Risk flag: review. Note focus: That the strongest primary legal record already separates the question of model training from the question of how the training corpus was assembled.
- Appendix A5. Companion note: Note N-029. Risk flag: review. Note focus: That large language models can sometimes reproduce identifiable verbatim material from their training data rather than only abstract statistical patterns.
- Appendix A6. Companion note: Note N-030. Risk flag: review. Note focus: That there is serious academic legal scholarship directly challenging the idea that generative AI training automatically falls under familiar exceptions.

- Appendix A10. Companion note: Note N-031. Risk flag: review. Note focus: That the dossier’s copyright / weights / compensation framing entered a bipartisan U.S. congressional policy record by December 17, 2024.
- Appendix A9. Companion note: Note N-032. Risk flag: low. Note focus: That AI systems depend on significant human labor for data generation, annotation, verification, and operational correction.
- Appendix A8. Companion note: Note N-033. Risk flag: medium. Note focus: That another federal court outside the Anthropic case treated training-stage copying and dataset assembly as central issues serious enough to survive dismissal.
- Appendix A2. Companion note: Note N-034. Risk flag: low. Note focus: That a major OpenAI training corpus was explicitly built from socially filtered outbound links shared on Reddit.
- Appendix A11. Companion note: Note N-035. Risk flag: medium. Note focus: That the dossier has a coherent thesis about how weights condense previously captured human labor.
- Appendix A4. Companion note: Note N-037. Risk flag: review. Note focus: That by August 2023 the U.S. Copyright Office had already elevated AI training inputs and disclosure questions into a formal federal inquiry.
- Appendix A1. Companion note: Note N-038. Risk flag: review. Note focus: That the U.S. Copyright Office treats AI training on copyrighted works as a major live legal and policy problem rather than a settled matter.