DOES DIFFERENTIAL PRIVACY SOLVE COPYRIGHT?

A high-level discussion of copyright, privacy, and AI

PRESENTED BY

Tino Trangia

AGENDA

01 **WHAT IS COPYRIGHT?**

02 **RESEARCH OVERVIEW**

03 **IMPLICATIONS AND TAKEAWAYS**

WHAT IS COPYRIGHT?

COPYRIGHT: A BRIEF PRIMER

A type of intellectual property that gives its owner the **exclusive legal right to copy, distribute, adapt, display, and perform a creative work**, usually for a limited time.

What is the intent?

“To promote the Progress of Science and useful Arts…” (U.S. Const. art. I, § 8, cl. 8)

Key concepts

- Protects expression, not ideas
- Copying = access + substantial similarity

Fair use doctrine

- Purpose and character of use (transformative vs. substitutive)
- Nature of copyrighted work
- Amount and substantiality of the portion used
- Effect of the use upon the potential market

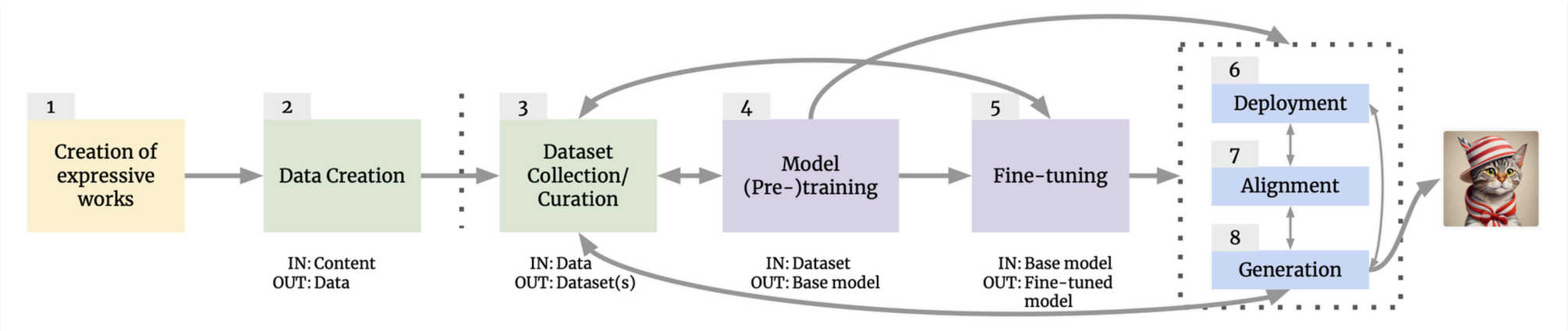
GENERATIVE AI IS UPENDING 200 YEARS OF U.S. COPYRIGHT LAW

New challenges

- AI ingests and outputs data at an unprecedented scale
- Humans input ideas and questions (not protected), model outputs expression (not protected?)
- Substantial similarity becomes difficult to prove
- Models don’t need incentives to produce new works → beyond the scope of copyright law?

Infringement can occur in training or deployment

- When does copying work for training constitute infringement?
- When are AI-generated outputs considered derivative (substantially similar) works?



(Lee et al, 2024)

RESEARCH OVERVIEW

Extracting Training Data from Large Language Models (Carlini et al, 2021)

GPT-2 memorizes and can be prompted to regurgitate verbatim training data

Extracting Training Data from Diffusion Models (Carlini et al, 2023)

Diffusion models memorize and reproduce training data exactly

Quantifying Memorization Across Neural Language Models (Carlini et al, 2022)

Memorization scales with model size and training data duplication

Foundation Models and Fair Use (Henderson et al, 2023)

Discusses whether training foundation models falls into fair use doctrine

Provable Copyright Protection for Generative Models (Vyas et al, 2023)

Introduces near access-freeness, the first mathematical framework for provable copyright protection

Can Copyright be Reduced to Privacy? (Elkin-Koren et al, 2023)

Argues that algorithmic stability approaches are not aligned with the legal definitions and intent of copyright

Scalable Extraction of Training Data from (Production) Language Models (Nasr et al, 2023)

Demonstrates practical extraction attacks of memorized data from aligned production models

The Files are in the Computer: On Copyright, Memorization, and Generative AI (Cooper & Grimmelmanm, 2024)

Memorization is neither necessary nor sufficient for copyright infringement

Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain (Lee et al, 2024)

Analyzes copyright issues at all stages of gen AI (data → modeling → deployment), says NAF contradicts legal principles of expression source

How much do language models memorize? (Morris et al, 2025)

New metrics for measuring memorization and mitigation techniques.
3.6 bits per parameter capacity.

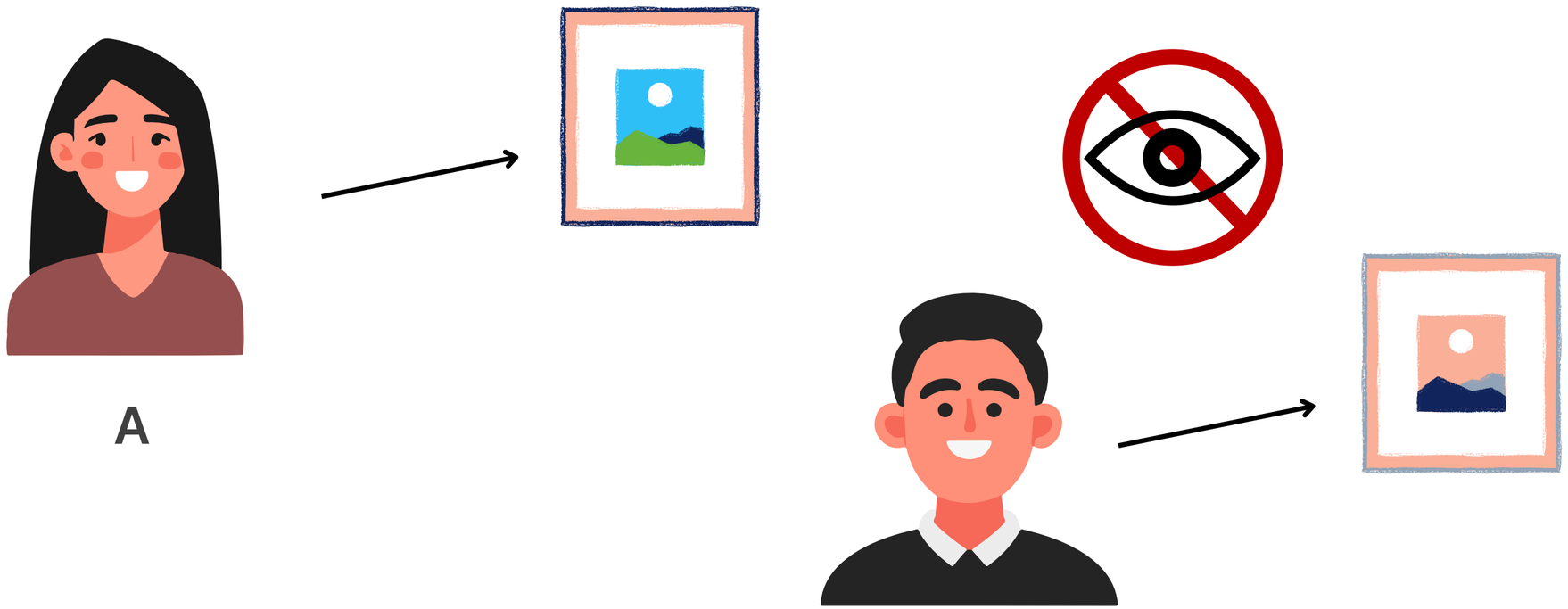
What my privacy papers (don't) have to say about copyright and generative AI (Carlini, 2025)

Carlini’s memorization papers are being misused in the courtroom: they only show that models sometimes output verbatim data

Blameless Users in a Clean Room: Defining Copyright Protection for Generative Models (Cohen, 2025)

Reveals issues with NAF and proposes a DP-based clean-room framework

HOW DOES DIFFERENTIAL PRIVACY APPLY TO COPYRIGHT?



A

B

CAN COPYRIGHT BE REDUCED TO PRIVACY? (ELKIN-KOREN ET AL, 2023)

The algorithmic stability angle

- DP interpreted as a criterion for originality
  - Painter's example: If Painter B never observed Painter A's work, would she still create the same art?
- DP → NAF (near access-freeness)

PROVABLE COPYRIGHT PROTECTION FOR GENERATIVE MODELS (VYAS ET AL, 2023)

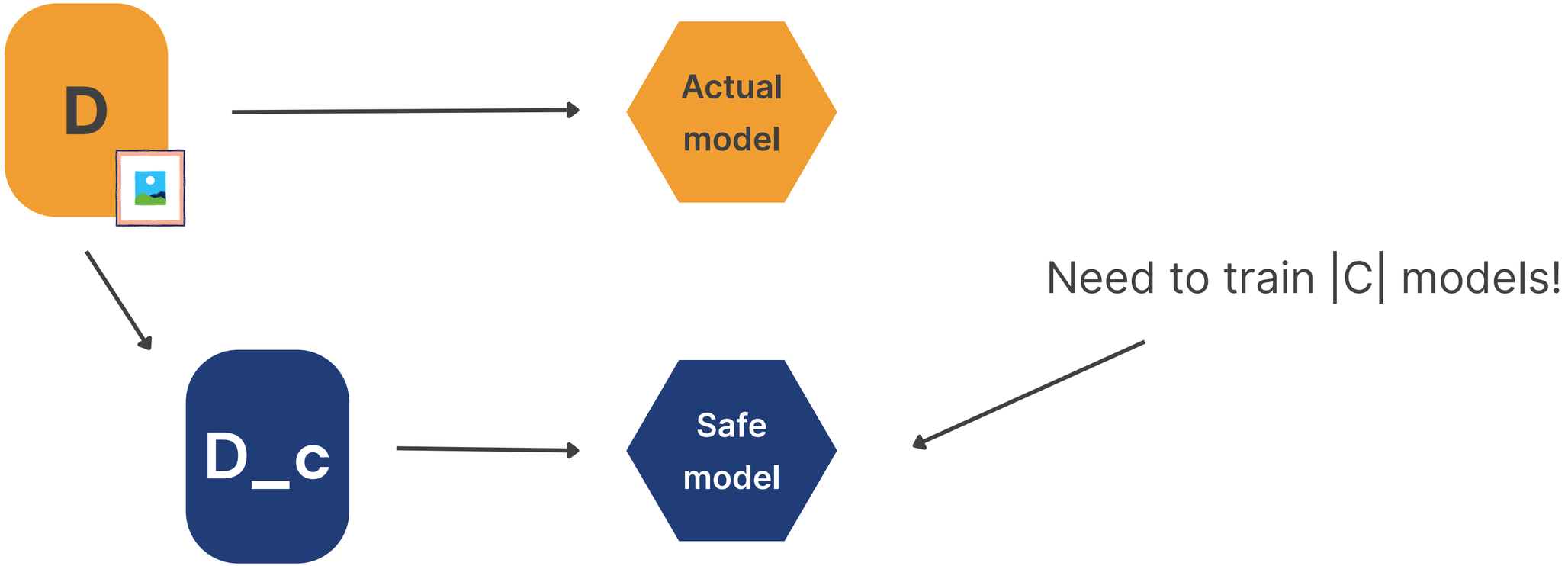
Near Access-Freeness (NAF)

- A model is k-NAF if its output is at most k-bits different from a “safe” model that never accessed the copyrighted content
- Quantitative question of the acceptable value of k
- Qualitative question of providing a “safe” function
- Essentially a relaxed variant of DP
  - Separates access and substantial similarity
  - One-sided (upper) bound
  - Only outputs matter → black-box algorithm

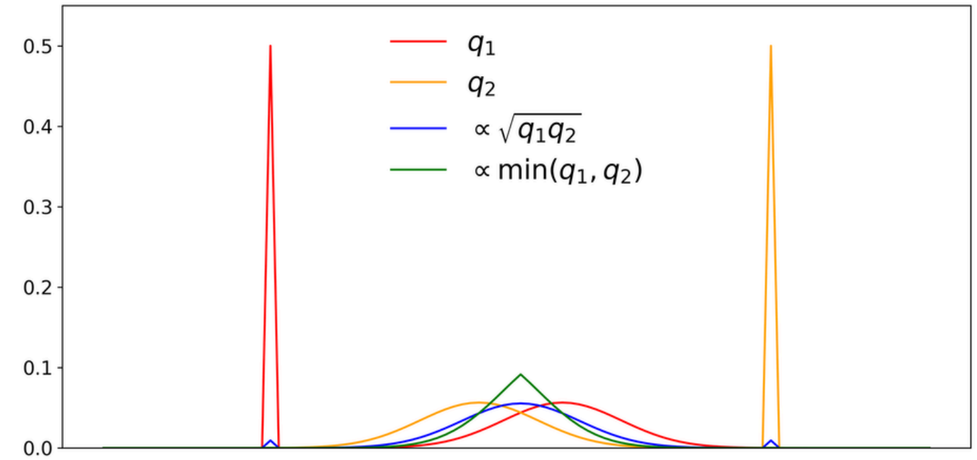
$$\Delta\Big(p(\cdot|x) \;\Big\|\; \mathsf{safe}_C(\cdot|x) \Big) \leq k_x$$

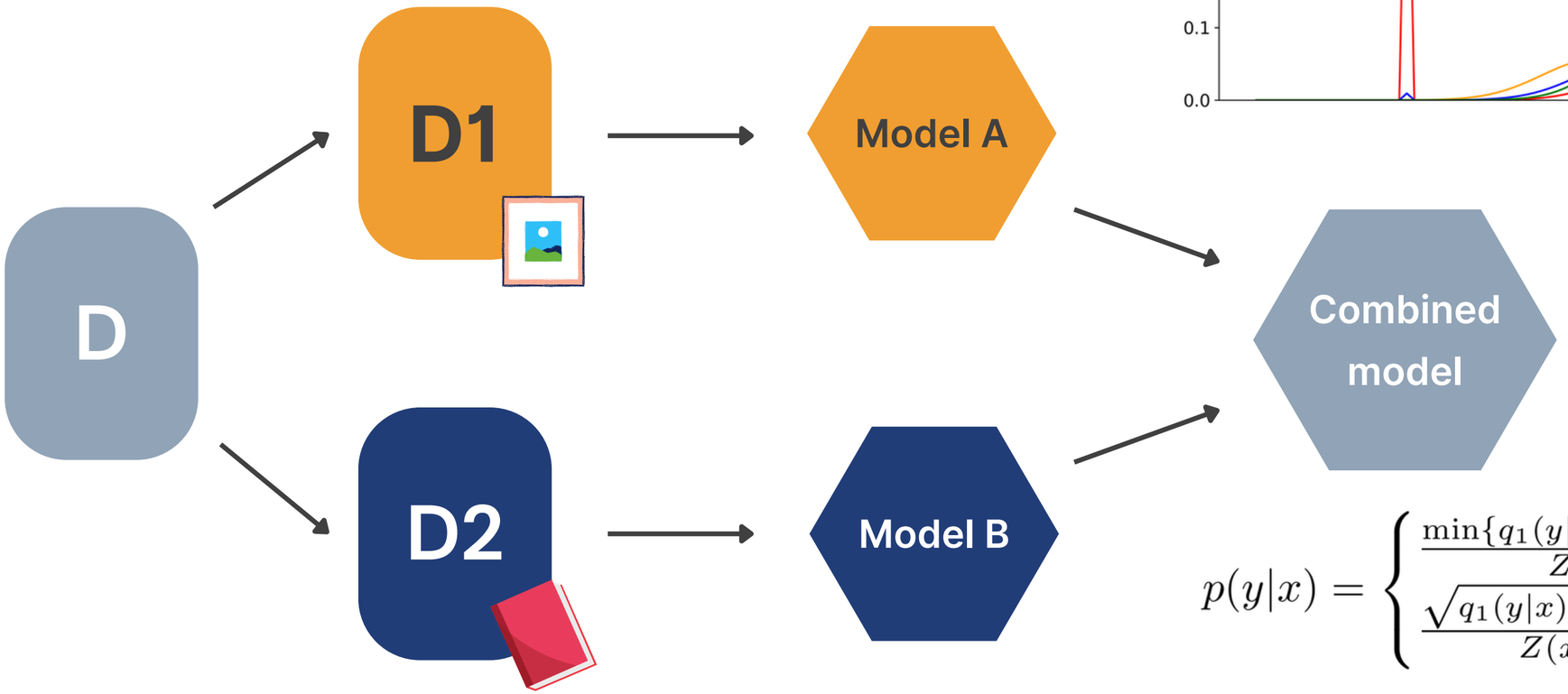
p is k-NAF if the above holds for all $x \in X$ with $k_x \leq k$

LEAVE ONE OUT



COPY-PROTECTION-Δ





$$p(y|x) = \begin{cases} \frac{\min\{q_1(y|x), q_2(y|x)\}}{Z(x)} & \text{if } \Delta = \Delta_{\max} \\ \frac{\sqrt{q_1(y|x) \cdot q_2(y|x)}}{Z(x)} & \text{if } \Delta = \Delta_{\mathrm{KL}}. \end{cases}$$

CAN COPYRIGHT BE REDUCED TO PRIVACY? (ELKIN-KOREN ET AL, 2023)

Copyright law is complicated

- Quantitative measures must align with a utilitarian rationale
- “Breathing room” for subsequent authors
- Privacy protects **content**, but copyright protects (original) **expression**
- Scope can vary among different elements of the same work

Privacy is fundamentally misaligned

- **Over-inclusive:** Withholds too much content from subsequent authors (de minimis, fair use, unprotected aspects)
- **Under-inclusive:** Can allow unlawful access to copyrighted expression, which need not come from the input content itself

CAN COPYRIGHT BE REDUCED TO PRIVACY? (ELKIN-KOREN ET AL, 2023)

- NAF circumvents some challenges with using DP in copyright
  - Content-safety vs. model-safety
  - Soft vs. hard constraints
  - Generalized safety functions
- Allows algorithms to be influenced by input content
- But there is no free lunch. Choose 2:
  - Allow learning from copyrighted works (safe models are influenced)
  - Protect the set of copyrighted works
  - Maintain model quality

BLAMELESS USERS IN A CLEAN ROOM: DEFINING COPYRIGHT PROTECTION FOR GENERATIVE MODELS (COHEN, 2025)

NAF is not safe

- Vulnerable to multiple prompt composition and data-dependent prompts
- NAF models can enable verbatim copying even if users know nothing about the underlying data (tainted model)

BLAMELESS USERS IN A CLEAN ROOM: DEFINING COPYRIGHT PROTECTION FOR GENERATIVE MODELS (COHEN, 2025)

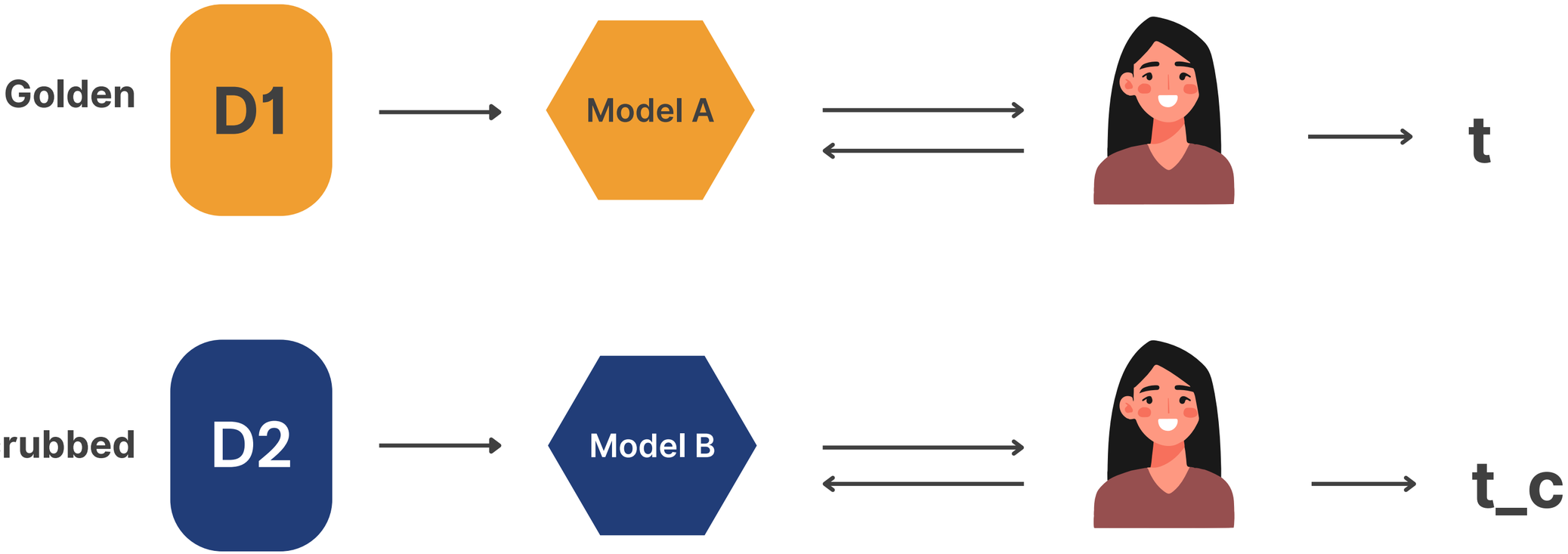
Clean-room copy protection

- Copyright dependency graph
- Scrub function creates sub-dataset that contains no works stemming from a particular copyrighted item

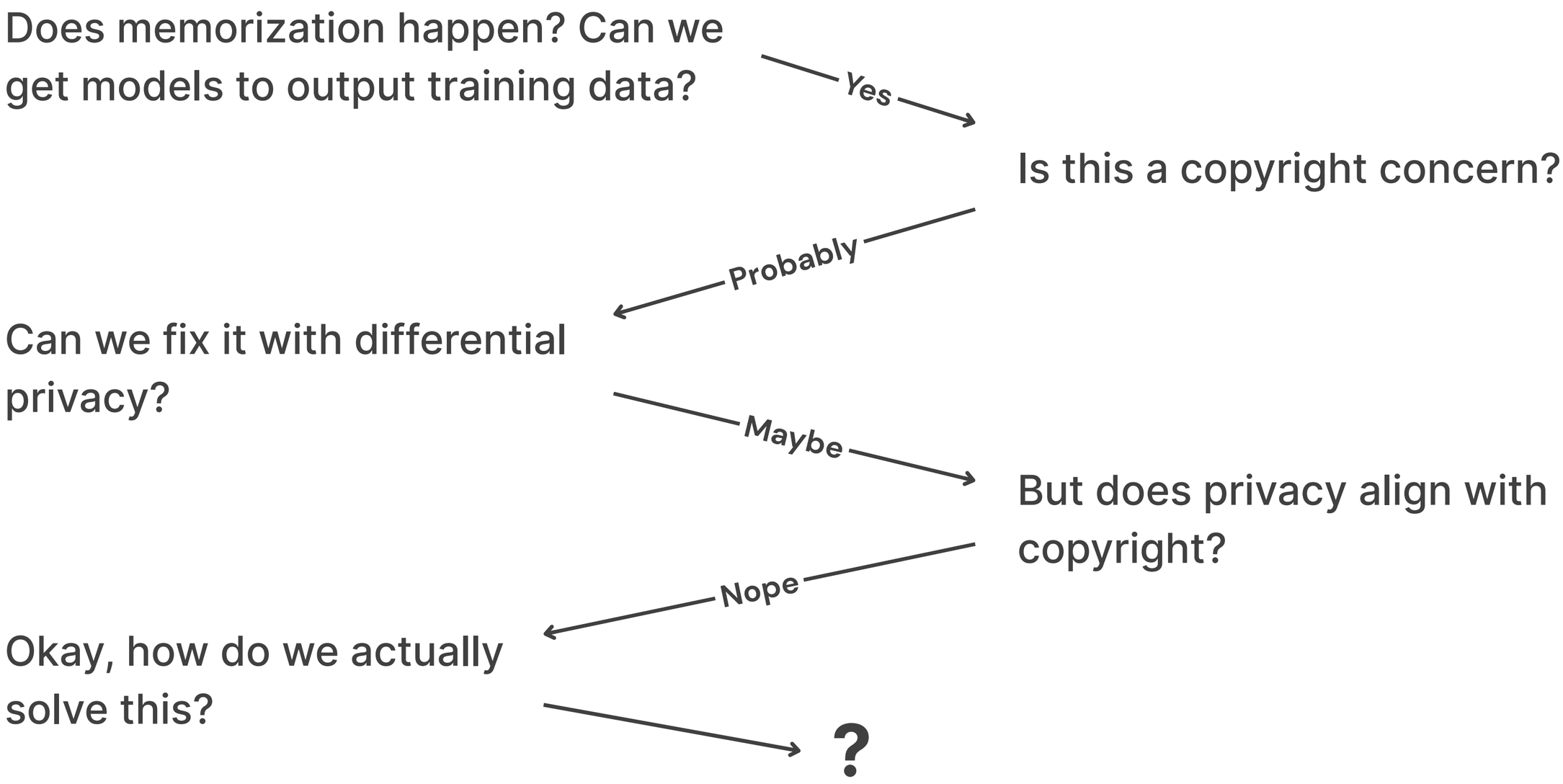
DP training implies clean-room protection

- Only if dataset is “golden” (copyright deduplication), no two items (potentially a derivative) stem from the same copyrighted work
- Protects the (honest) generative AI user from liability
  - Indemnifies user if data was not “golden“

- Had we scrubbed the data, the user would not have copied
- User's output distribution is similar in the real world → probability of producing substantially similar work is upper bounded
- **Sufficient** to prevent copyright infringement (solves **under**-inclusiveness)

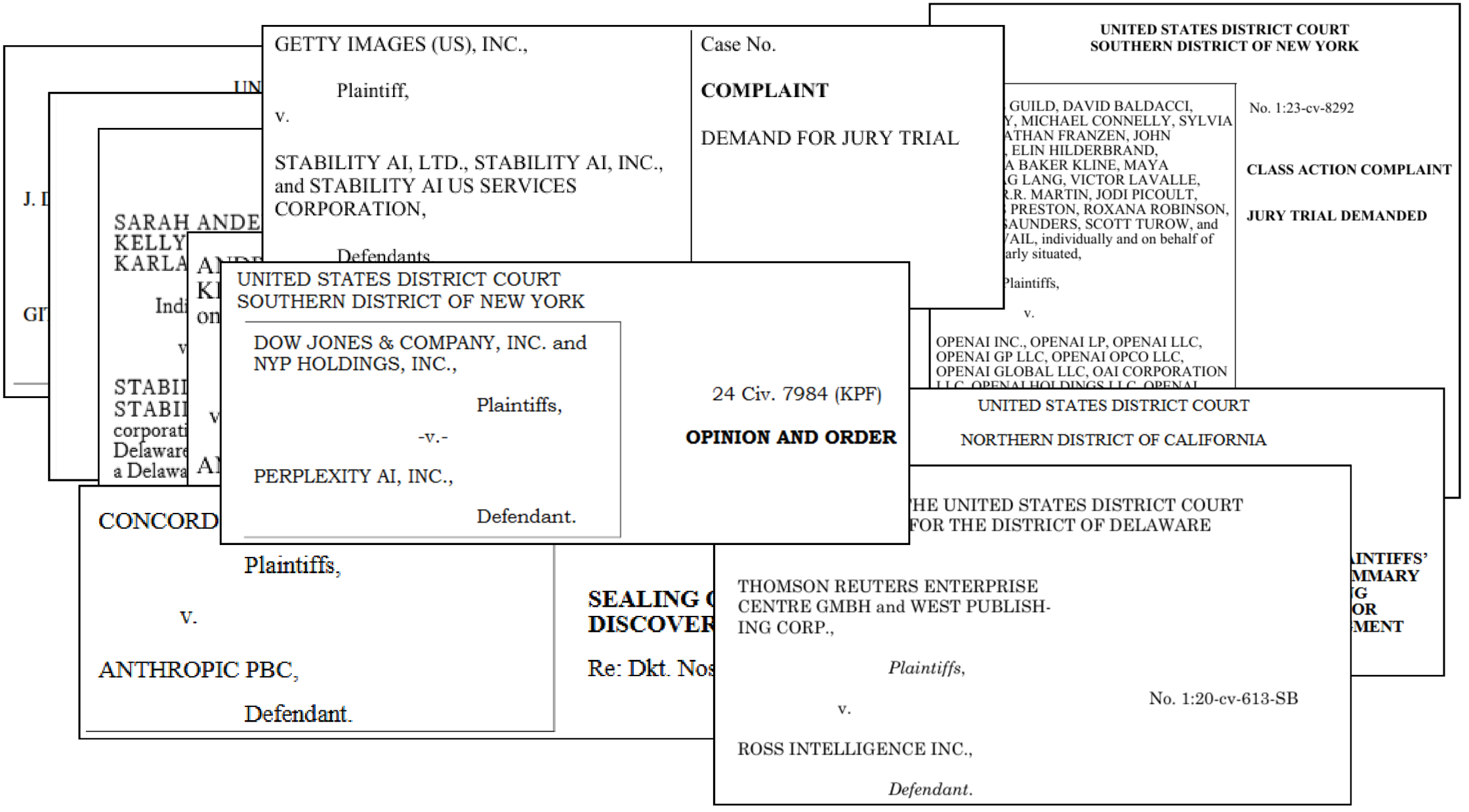


IMPLICATIONS AND TAKEAWAYS


Does memorization happen? Can we get models to output training data?
Yes
Is this a copyright concern?
Probably
Can we fix it with differential privacy?
Maybe
But does privacy align with copyright?
Nope
Okay, how do we actually solve this?
?

MEMORIZATION ≠ COPYRIGHT INFINGEMENT

PRIVACY ≠ COPYRIGHT PROTECTION



GETTY IMAGES (US), INC.,

Plaintiff,

v.

STABILITY AI, LTD., STABILITY AI, INC., and STABILITY AI US SERVICES CORPORATION,

Defendants

Case No.

COMPLAINT

DEMAND FOR JURY TRIAL

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

GUILD, DAVID BALDACCI, Y, MICHAEL CONNELLY, SYLVIA ATHAN FRANZEN, JOHN , ELIN HILDERBRAND, A BAKER KLINE, MAYA G LANG, VICTOR LAVALLE, R.R. MARTIN, JODI PICOULT, PRESTON, ROXANA ROBINSON, SAUNDERS, SCOTT TUROW, and /AIL, individually and on behalf of arly situated,

Plaintiffs,

v.

OPENAI INC., OPENAI LP, OPENAI LLC, OPENAI GP LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION LLC, OPENAI HOLDINGS LLC, OPENAI

No. 1:23-cv-8292

CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

DOW JONES & COMPANY, INC. and NYP HOLDINGS, INC.,

Plaintiffs,

-v.-

PERPLEXITY AI, INC.,

Defendant.

24 Civ. 7984 (KPF)

OPINION AND ORDER

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA

CONCORD

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

SEALING C
DISCOVER

Re: Dkt. Nos

HE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

THOMSON REUTERS ENTERPRISE CENTRE GMBH and WEST PUBLISHING CORP.,

Plaintiffs,

v.

ROSS INTELLIGENCE INC.,

Defendant.

No. 1:20-cv-613-SB

INTIFFS' MMARY G OR MENT

IMPLICATIONS

Technical

- Open questions in **mitigation, remedy, and compensation**
- Need new metrics and techniques oriented specifically for copyright
- Can’t use DP (by itself) to quantify infringement
- Machine unlearning
- Retrieval augmented generation (RAG) protections
- Provenance tracking
- Copyright takedowns

Judicial

- Applying fair use doctrine
  - Transformativeness
  - Market dilution
- What is permissible under current laws?
- How do we attribute harms?

Legislative

- Copyright reform?
- Licensing regulations, TDM exceptions
- Fair use codification, AI privacy laws
- Consent, compensation, control
- How do we balance competitiveness with protections?

KEY TAKEAWAYS

- Existing technical approaches predominantly address memorization, not copyright
- Privacy measures are both under- and over-inclusive: Copyright infringement can occur even without exact reproduction, and yet exact reproduction does not always violate copyright
- Other issues remain unaddressed (training data sourcing, market dilution)
- **Interdisciplinary collaboration is essential**

REFERENCES

1. N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, “Extracting training data from large language models,” 2021. [Online]. Available: https://arxiv.org/abs/2012.07805
2. N. Vyas, S. Kakade, and B. Barak, “On provable copyright protection for generative models,” 2023. [Online]. Available: https://arxiv.org/abs/2302.10870
3. P. Henderson, X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang, “Foundation models and fair use,” 2023. [Online]. Available: https://arxiv.org/abs/2303.15715
4. M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tram`er, and K. Lee, “Scalable extraction of training data from (production) language models,” 2023. [Online]. Available: https://arxiv.org/abs/2311.17035
5. A. Cohen, “Blameless users in a clean room: Defining copyright protection for generative models,” 2025 [Online]. Available: https://arxiv.org/abs/2506.19881
6. Lemley, Mark A., How Generative AI Turns Copyright Upside Down (July 21, 2023). Available at SSRN: https://ssrn.com/abstract=4517702 or http://dx.doi.org/10.2139/ssrn.4517702
7. N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tram`er, B. Balle, D. Ippolito, and E. Wallace, “Extracting training data from diffusion models,” 2023. [Online]. Available: https://arxiv.org/abs/2301.13188
8. Lee, Katherine and Cooper, A. Feder and Grimmelmann, James, Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain (July 27, 2023). Forthcoming, Journal of the Copyright Society 2024, Available at SSRN: https://ssrn.com/abstract=4523551 or http://dx.doi.org/10.2139/ssrn.4523551
9. N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tram`er, and C. Zhang, “Quantifying memorization across neural language models,” ArXiv, vol. abs/2202.07646, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246863735
10. J. X. Morris, C. Sitawarin, C. Guo, N. Kokhlikyan, G. E. Suh, A. M. Rush, K. Chaudhuri, and S. Mahloujifar, “How much do language models memorize?” 2025. [Online]. Available: https://arxiv.org/abs/2505.24832
11. O. Bousquet, R. Livni, and S. Moran, “Synthetic data generators: Sequential and private,” 2020. [Online]. Available: https://arxiv.org/abs/1902.03468
12. Cooper, A. Feder and Grimmelmann, James, The Files are in the Computer: On Copyright, Memorization, and Generative AI (April 22, 2024). Cornell Legal Studies Research Paper, Chicago-Kent Law Review, Volume 100, pp. 141-219 (2025), Available at SSRN: https://ssrn.com/abstract=4803118

THANK YOU

Questions, comments, concerns?