

The Legal Life of a Dataset

When training data is legally usable, and what counsel needs
to know

Janel Thamkul, General Counsel at Fireworks AI and Visiting Senior Fellow at the Institute for Law & AI

October 9, 2026

NOT LEGAL ADVICE - CONSULT A LAWYER :)

The request usually looks like this

MONDAY, 9:14AM

quick question: we found a corpus of 2 million posts from a public health discussion site.
can we train on it? hoping to start the run Thursday

What was asked

A yes or no by Thursday

What was assumed

Public means available. Available means usable.

What is missing

The facts that determine the answer

The answer is a spectrum and usually depends on scope

Which records? For what use? In what kind of model? Released how? In which markets? Good legal advice should explain the conditions for a yes.

The intake

01

Where did it come from?

Source and provenance

02

How did we get it?

Acquisition method

03

What commitments apply?

Terms accepted and promises made

04

What is in it?

Content categories

05

How will we use it?

Training stage and application

06

How will we release it?

Release method and markets

01 Where did it come from?

What counsel needs to know

- Can you identify the source of each record, or at least each category of records?
- Is there a documented chain from the original publisher to your copy?
- Did you record that chain when the data was collected, or reconstruct it later?
- Do you know which terms applied when the data was collected?

Answers that slow the review

- "It was on a shared drive"
- "We pulled it from a mirror"
- "It was already in the pretraining mix"
- "I think someone licensed it"
- "Someone who left got it from a friend"
- "It fell off the back of a truck"

Without provenance, the rest of the analysis breaks down. Unknown lineage makes a corpus hard to approve, filter, defend, or delete.

02 How did we get it?

The same data can carry different legal risk depending on how it was acquired.

Compliant crawl

Site controls respected and rate limits honored

Controls bypassed

A login wall, rate limit, or block was worked around

Licensed

Negotiated terms define the scope and obligations

Collected from users

Collected through your product and governed by your own promises

Aggregator or mirror

You inherit someone else's collection choices and risk

Unauthorized copy

For example, a shadow library or leaked dataset. Often disqualifying.

03 What commitments apply?

Terms we accepted

- Site terms of service
- API terms and developer agreements
- Dataset licenses
- Data purchase and sharing agreements
- Another model provider's output terms

Promises we made

- Our privacy policy at the time
- Our user agreement and product promises
- Enterprise commitments and DPAs
- Public statements about data use
- Consent language shown at signup

Contract can control even when copyright or privacy law would permit the use. A favorable fair use position does not override a term you accepted.

04 What is in it?

Personal data

Identifiable, directly or in combination

Special category data

Health, biometric, religious, political, union membership, or sexual orientation

Children's data

Subject to separate consent and design rules in many markets

Regulated records

Health, financial and education records under sector rules

Confidential material

Trade secrets and documents under a duty of confidence

Copyrighted works

Books, news, film and code under restrictive licenses

Per se unlawful

Content that is illegal to possess

Unknown

The honest answer for many web-scale corpora

Calling it a “public web corpus” identifies the source, not the contents.

05 How will we use it?

Training stage

Pretraining, fine-tuning, alignment, retrieval, or evaluation. Scale and memorization risk vary across these uses.

Application

A general assistant and a clinical decision support tool face different rules and different consequences for error.

Decisions about people

If outputs inform decisions about people, additional rules may apply, including automated-decision, sector-specific, and high-risk AI rules.

This is where the team has room to narrow the request. A limited use can often turn a no into a yes.

06 How will we release it?

Served through an API

- You retain an output-filtering layer
- You can patch the model, log activity, or revoke access
- You can respond to deletion and takedown requests
- Contracts can limit markets and uses

Released as open weights

- No downstream control or recall
- Memorized content is hard to contain
- More downstream uses are foreseeable
- Export and distribution rules apply

The same dataset may be approvable for an API deployment but not an open release. Release method and system mitigations can matter more than the data itself.

Provenance and contract are often the real constraints

Statutes (e.g., DMCA, GDPR) get the headlines. In practice, an undocumented source or an old contract term may be what stops the run.

Four legal regimes, four different questions

Legal regime	What it protects	Where it applies	Where roadblocks arise
Copyright	Protected expression	Copies and outputs	No exception covers the use in the relevant market
Contract and terms	Promises you accepted or made	Acquisition and permitted use	A term prohibits the use even if the law allows it
Privacy	Personal data	Collection and purpose	No lawful basis applies
Sector rules	Certain types of records	The record wherever it goes	The rule continues to apply to the record

These regimes are cumulative. Clearing one does not clear the others, and one dataset may trigger all four.

Copyright requires three separate analyses

STAGE ONE

Was the copy lawful?

How you obtained and copied the material is assessed separately. An unlawfully acquired copy creates its own risk.

STAGE TWO

Does an exception apply?

In the U.S., this usually means fair use, with transformativeness and market effect carrying significant weight. Elsewhere, the use needs a specific exception or permission.

STAGE THREE

What does the model output?

Do outputs reproduce protected expression or substitute for the original work?

A favorable answer at stage two does not necessarily cure stage one or resolve stage three.

The same dataset across six markets

Market	Route for training copies	Key limitation
United States	Fair use, case by case	Case-specific; acquisition analyzed separately
European Union	Text and data mining exceptions	Rightsholders may reserve works from commercial mining
Japan	Broad statutory exception	Unavailable if use unreasonably harms rightsholders
Singapore	Computational data analysis	Lawful access required; contracts cannot override
United Kingdom	Noncommercial research	No general commercial exception
Australia	Fair dealing; listed purposes only	No mining exception for this use

A dataset may be usable in one market and unusable in another. Training location, release markets, and customers each matter.

Four regimes

Privacy law applies at collection and to later uses

Under GDPR

Collection and storage are processing. You need a lawful basis when the data is first collected, before training begins. A privacy-preserving training method does not cure a problem at collection.

In the United States

There is no general federal lawful-basis requirement. Instead, sector laws, state privacy and biometric laws, and the FTC regulate how data is obtained, described, and used.

Why that matters

Techniques that reduce memorization address a real risk, but not the legal basis for collecting and using the data.

Legitimate interest is usually the key privacy argument

A legitimate interest

Usually straightforward. Developing a model can qualify as a legitimate interest.

Necessity

Could you meet the goal with less data or different data? A training technique does not answer that.

The balance

Weigh the interest against people's rights and reasonable expectations.

Where mitigations help

They can reduce the severity of the impact by limiting how much any person's data contributes.

Where mitigations may not help

They do not change reasonable expectations. A person posting to a forum in 2015 might not have expected model training.

Regulators treat safeguards as relevant to the balancing test, but still assess each case individually. They also do not assume that trained models are anonymous.

Special-category data has no easy route

Additional legal basis required

The special-category restriction sits on top of the ordinary lawful-basis requirement. Legitimate interest alone is not enough.

Made manifestly public

Requires a deliberate act by the person. Public availability alone is not enough.

Research exception

Requires a basis in EU or Member State law. The rules are fragmented and often do not fit commercial research.

Indirect revelation

Data can be in scope if it reveals special-category information, even when nobody intended to collect it.

Every web-scale corpus is likely to contain some data for which no legal basis is available. Teams also lack reliable ways to find it, and downstream training techniques do not address it.

Some rules follow the data wherever it goes

Health records

Protected health information has two defined de-identification paths: remove a specified list of identifiers or obtain an expert determination of low re-identification risk.

Financial and education

Separate consent, notice, and disclosure rules apply to these records wherever the data ends up.

Biometric laws

State laws may require consent and allow private lawsuits or statutory damages without proof of harm.

Why these rules can be easier to apply

Counsel gets a checklist or a defined expert standard. That makes the rule easier to apply. A weaker protection with a clear standard may clear faster than a stronger protection without one.

Legal gates and technical mitigations

LEGAL GATES

Lawful basis, terms accepted, the copy itself

Purpose limits, record category, retention

Whether the use is excused

Release, erasure duties, distribution

Reproduction, substitution, regulated advice

PIPELINE

Acquisition

Corpus

Training

Weights

Outputs

TECHNIQUES

Source allowlists, crawler compliance

Detection, redaction, deduplication

Differential privacy, canaries

Memorization evals, unlearning

Filters, grounding, attribution

The strictest legal gates appear early in the pipeline. Most technical mitigations operate later.

What each technique can and cannot address

Technique	Risk it reduces	What it does not address
Differential privacy	Individual influence and memorization	Lawful basis at collection or contractual restrictions
Deduplication	Measurable verbatim reproduction and leakage	How the material was acquired
Identifier redaction	Direct identification from the corpus	Sensitive attributes inferred from remaining data
Federated learning	Movement and central storage of data	Purpose, lawful basis, or prior promises
Output filtering	Risk at the point of generation	Anything upstream of the model

These techniques reduce real risk, and counsel will rely on them. They cannot supply a missing lawful basis or override a contract term.

Synthetic data still implicates all four regimes

Privacy

Creating synthetic data from personal data is itself processing. The output may also remain personal data if someone can still be identified.

Sector rules

Health data offers an expert-determination path. The identifier checklist does not translate well to generated records.

Copyright

Questions remain about the output, what it derives from, and who owns the rights.

Contract

The source model's terms may prohibit using outputs for training. Teams often check this last, even when it controls the answer.

Synthetic data remains unsettled across all four regimes. The more utility it preserves, the harder it becomes to argue that it carries nothing from the source.

CONCLUSION

Technical guarantees and legal standards need a shared vocabulary

Translation between technical mitigations and legal implications is critical. Your counsel is your partner!

Janel Thamkul