PRO.YOUR LIFE

Reading a Dataset

How to tell whether a model was trained on work it had no right to use.

PRO.YOUR LIFE, INC. · A NONPROFIT CREATIVE TRUST · PROYOURLIFE.ORG FREE RESOURCE · © 2026

INTRODUCTION

What was it trained on?

How to tell whether an AI model was trained on work it had no right to use.

Every generative AI model learns from a dataset: a large collection of text, images, audio, or video. What is in that dataset shapes what the model can do, and whether it was built fairly. If creators' work was taken without consent, the model carries that with it.

Most companies do not make this easy to see. This explainer gives you the vocabulary, the places to look, and the questions to ask, so you can make informed choices about the tools you use and the platforms you share with.

What you will learn

- What a training dataset is, in plain language
- Where to find information about a model's training data
- How to decode the vague phrases companies use
- Green flags and red flags
- What to do if you think your work was used

IN THIS SERIES

Reading a Dataset is Volume 03 of the Pro.Your Life Ethical AI Library. Keep going with **Ethical AI Principles** (Volume 01) and **AI as a Creative Tool** (Volume 02), free at proyourlife.org.

ONE

Datasets, in plain language

A training dataset is the material a model studies to learn patterns. A music model studies recordings. An image model studies pictures paired with descriptions. A language model studies text. The model does not usually store each item like a library, but it is shaped by all of them, and in some cases it can reproduce parts of what it saw.

Datasets are often built by collecting material from across the web at huge scale, sometimes billions of items. Some large public datasets built this way have been at the center of lawsuits over whether creators' work was used without permission. Other datasets are built from licensed collections, public domain material, or content people chose to contribute.

SOURCE TYPE	WHAT IT MEANS	WHAT TO KNOW
Web scraped	Collected automatically from websites.	Often includes copyrighted work used without consent.
Licensed	Paid for or permitted by the rights holders.	Creators or their representatives agreed to the use.
Public domain	Material no longer under copyright.	Generally free to use.
Contributed	Given by users or creators on purpose.	Depends on whether the consent was informed and specific.
Synthetic	Generated by other AI models.	Inherits whatever was in the original model's data.

TWO

Where to look

Information about training data is scattered. These are the places it is most likely to be.

01 **Training data summaries.**
In the European Union, providers of general purpose AI models must publish a public summary of the content used to train them, using a standard template.

02 **Training data documentation.**
In California, developers of generative AI systems made available to Californians must post documentation about the data used to train them, including its sources and whether it includes copyrighted material.

03 **Model cards.**
Short documents that describe a model: what it is for, how it was built, its limits, and often its training data.

04 **Datasheets and dataset cards.**
Documents that describe a dataset itself: who made it, what is in it, how it was collected, and under what license.

05 **Terms of service and privacy policies.**
These tell you whether the company can use your uploads to train future models.

06 **Search tools.**
Some tools let creators search whether their work appears in certain public datasets. Coverage is limited, but a match is useful evidence.

THREE

Decoding the language

Companies often describe their data in phrases that sound reassuring and say very little.

WHEN THEY SAY	ASK THIS
“Publicly available data”	Anything that could be viewed online. Public is not the same as licensed. Your posted work is public.
“Data from the web”	Usually means scraped. Ask what was collected and whether opt outs were honored.
“Licensed and publicly available data”	A mix. Ask what share is licensed, and from whom.
“We respect opt outs”	Ask when opt outs started being honored, and whether earlier training is affected.
“Commercially safe”	Usually a promise about legal risk for the user, not about how creators were treated. Ask what the model was trained on.
“Trained on our own content”	Ask whose content that was originally, and whether contributors agreed to AI training.

FOUR

Green flags and red flags

GREEN FLAGS

- A public, specific list or summary of training data sources
- Training on licensed, public domain, or opt in material
- Creators paid or credited for training use
- Your uploads are not used for training by default
- Clear, plain language terms

RED FLAGS

- No information at all about training data
- Vague phrases like “publicly available” with no detail
- Outputs that closely imitate specific living artists
- Terms that claim rights to your uploads for training
- Opt out buried or impossible

FIVE

If you think your work was used

01 **Document it.**
Save screenshots, outputs that resemble your work, dates, and links. Keep your own drafts and files that prove your authorship.

02 **Search where you can.**
Use dataset search tools and the model's published documentation to look for your work or your name.

03 **Opt out going forward.**
Use platform settings and opt out tools to reduce future use, and update your website and contracts.

04 **Ask directly.**
Write to the company and ask whether your work was used and how to have it removed from future training.

05 **Get advice.**
Talk to a lawyer who works with creatives, or a creator rights organization, before taking legal action. The law in this area is changing quickly.

A dataset scorecard

- ☐ I can find information about this model's training data.
- ☐ The sources are specific, not vague.
- ☐ The data is licensed, public domain, or contributed with consent.
- ☐ My uploads are not used for training without my choice.
- ☐ The company answers creator questions directly.

Five checks is a tool worth considering. Two or fewer is a tool to question.

STAY IN THE WORK

Build, protect, and sustain your creative work.

Pro.Your Life is a nonprofit creative trust helping artists and creatives build, protect, and sustain their work in a technology driven world without losing ownership, culture, or humanity. We publish free educational resources, host workshops, and are building a global creative trust for the age of AI. Join the founding community at proyourlife.org.

This resource is education, not legal advice. Every contract and every situation is different. When real money or real rights are on the line, talk to a lawyer who works with creatives.

Pro.Your Life and the fader mark are trademarks of Pro.Your Life, Inc. (application pending). © 2026 Pro.Your Life, Inc. You may share this resource freely in its complete, unaltered form.