Position: Data Authenticity, Consent, and Provenance for AI Are All Broken

What will it take to fix them?


Shayne Longpre[1]


Robert Mahari[1]


Naana Obeng-Marnu[1,2]


William Brannon[1,2]

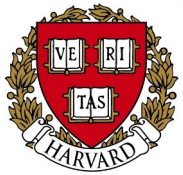

Tobin South[1]


Katy Ilonka Gero[3]


Alex Pentland[1]

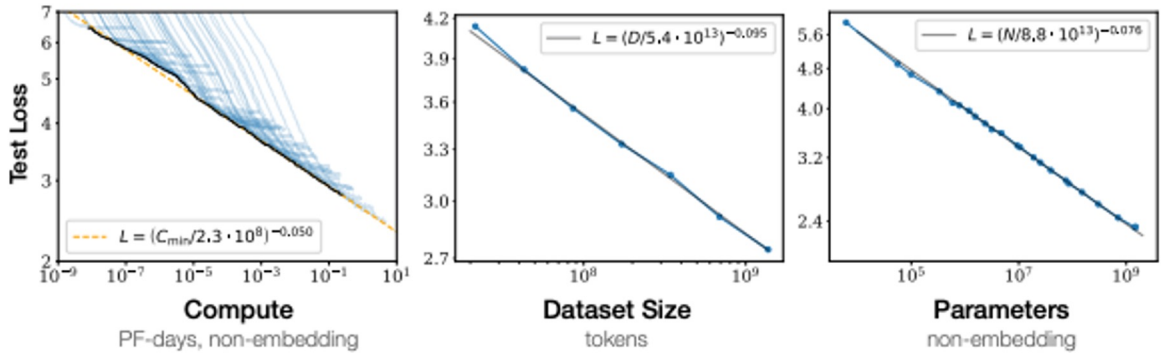

Jad Kabbara[1,2]

[1] MIT Media Lab
[2] MIT Center for Constructive Communication
[3] Harvard University



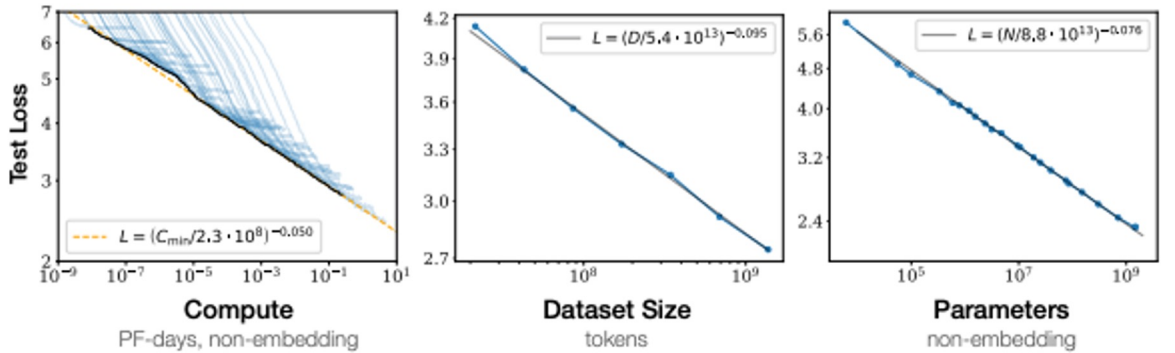
ICML
International Conference
On Machine Learning

What’s the problem?



Kaplan et al (2020)

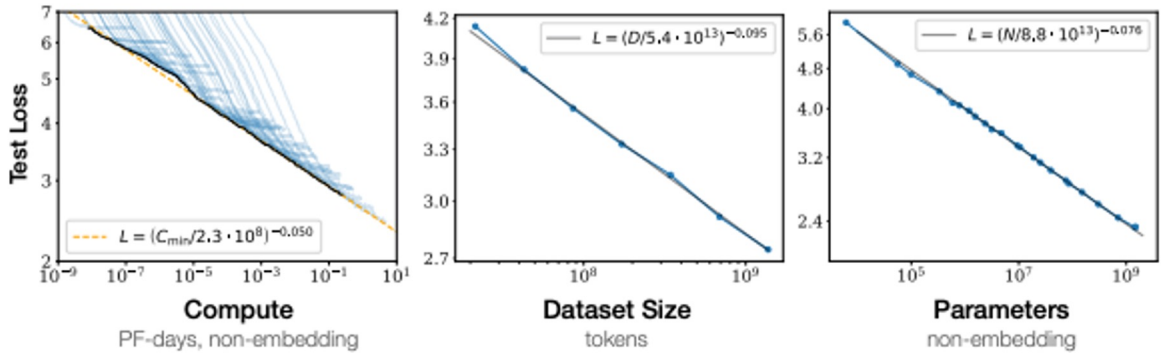
Generative AI needs huge scale!



Kaplan et al (2020)

Generative AI needs huge scale!

Including lots of data.

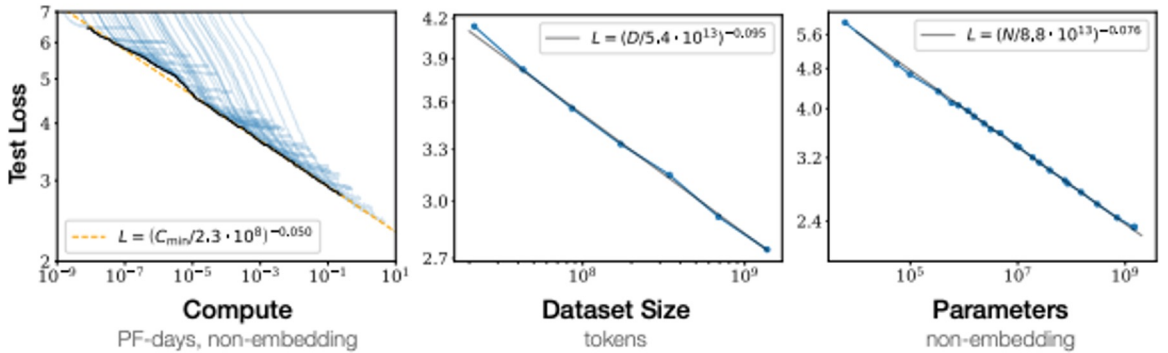


Kaplan et al (2020)

Generative AI needs huge scale!

Including lots of data.

But where does the data come from?



Kaplan et al (2020)

Generative AI needs huge scale!

Including lots of data.

But where does the data come from?

And what's in it?

We don’t really know!

Limited visibility into composition of commercial training data

Limited visibility into composition of commercial training data

Open models are more transparent, but underlying datasets are often not documented thoroughly or consistently (Longpre et al, 2023)

Why is it important?

Terms of use

OpenAI suspends ByteDance's account after it allegedly used GPT to build rival AI product: report

- New York Post, Dec 18

CSAM & illegal content

TECH / ARTIFICIAL INTELLIGENCE

AI image training dataset found to include child sexual abuse imagery

- The Verge, Dec 20

Copyright

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

- New York Times, Dec 27

Leads to several problems

Personal information (PII) leaks

ChatGPT Can Reveal Personal Information From Real People, Google Researchers Show

- Vice, Nov 29

Bias and discrimination

AI shows clear racial bias when used for job recruiting, new tests reveal

- Mashable, March 8

Leads to several problems

65%

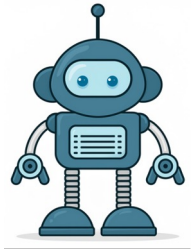
of HF datasets in a recent large-scale audit have incorrect licenses

Longpre et al (2023)

What’s been done?

Provenance standards:

- Datasheets, data statements, data cards, data nutrition labels, D&TA DPS
- Important! But unevenly adopted



. txt

robots.txt:

- Aimed at search crawlers, not AI
- Next gen: “learners.txt” or consent registries like SpawningAI

Watermarking (a la C2PA):

- Either natural or generated content
- Can often be defeated or removed
- Difficult to do for text

Images via Adobe Firefly

Common Crawl:

- Standard source of pretraining text data
- Limited metadata, mainly text, doesn't collect or break out fine-tuning data

Hugging Face Datasets:

- Very popular ML dataset host
- Metadata is crowdsourced and frequently incorrect

Data Provenance Initiative:

- Structured repository of human-collected metadata on ML datasets
- So far, only text fine-tuning data
- Human collection = less scalable

What’s missing?

Much provenance information simply isn’t collected.

Wide adoption

Much provenance information simply isn’t collected.

For example, unspecified license information on some datasets.

Wide adoption

Much provenance information simply isn’t collected.

For example, unspecified license information on some datasets.

Not having this info **impedes research**, causes **friction for industry** and **blinds regulators**.

Wide adoption

There are many ways to track provenance, authenticity and consent information…

Systematic adoption

There are many ways to track provenance, authenticity and consent information…

…which are less than the sum of their parts.

Systematic adoption

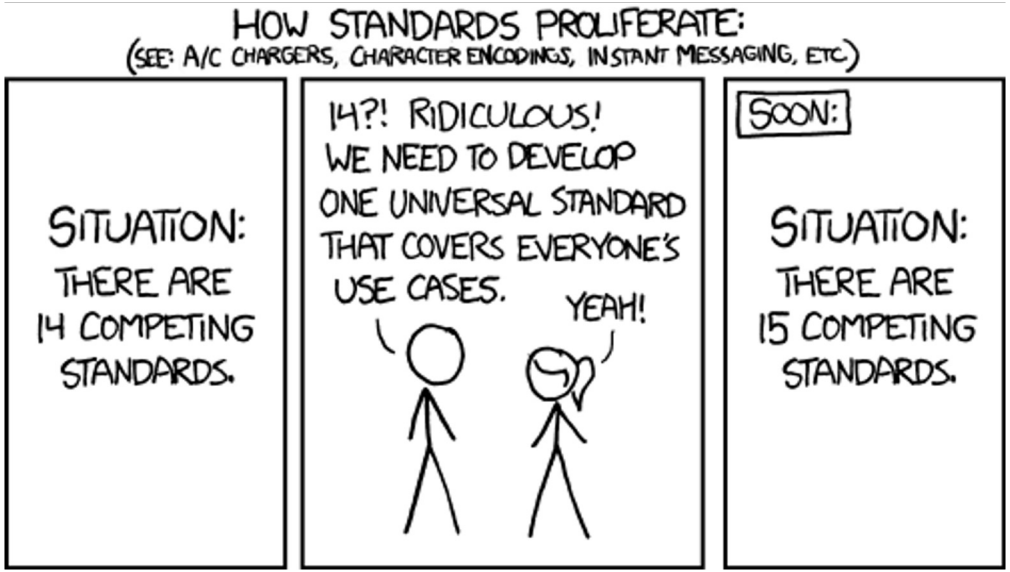
There are many ways to track provenance, authenticity and consent information…

…which are less than the sum of their parts.

Lack of standardization makes metadata **hard to use** and **hard to maintain**.

Systematic adoption

Recommendations



xkcd

We're not proposing a new standard

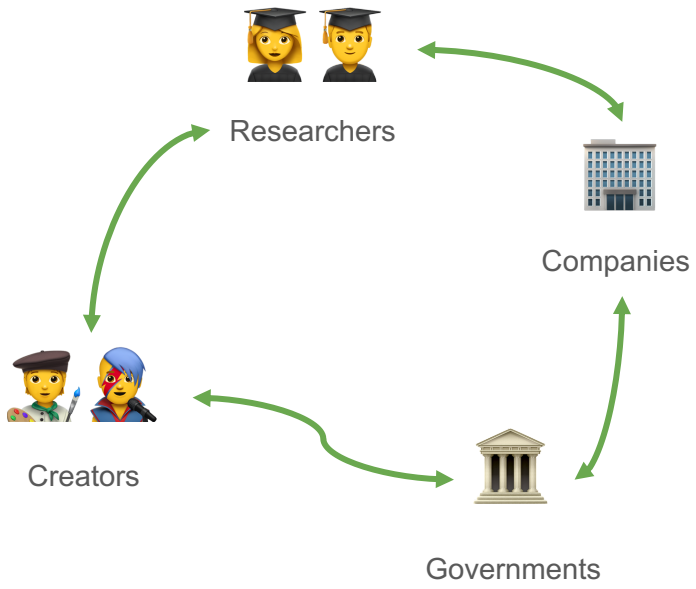
A holistic view is essential

This is one problem, with many parts:

- **Science**: More predictable and interpretable models
- **Law**: Reducing legal risk & aiding compliance
- **Policy**: Informing regulators
- **Equity**: Benefiting dataset creators

They fit together! Similar changes can help all.

This is a problem for the whole AI/ML community:



Community-wide problems need community-wide solutions

Check out the paper!

Thank you!

slongpre@media.mit.edu

wbrannon@media.mit.edu