Bridging the Provenance Gap across Text, Speech, & Video

Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, Nayan Saxena, Ahmad Mustafa Anis, Emad A. Alghamdi, Vu Minh Chien, Da Yin, Kun Qian, Yizhi Li, Minnie Liang, An Dinh, Shrestha Mohanty, Deividas Mataciunas, Tobin South, Jianguo Zhang, Ariel N. Lee, Campbell S. Lund, Christopher Klamm, Damien Sileo, Diganta Misra, Enrico Shippole, Kevin Klyman, Lester JV Miranda, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Vipul Gupta, Vivek Sharma, Xuhui Zhou, Caiming Xiong, Luis Villa, Stella Biderman, Alex Pentland, Sara Hooker, Jad Kabbara

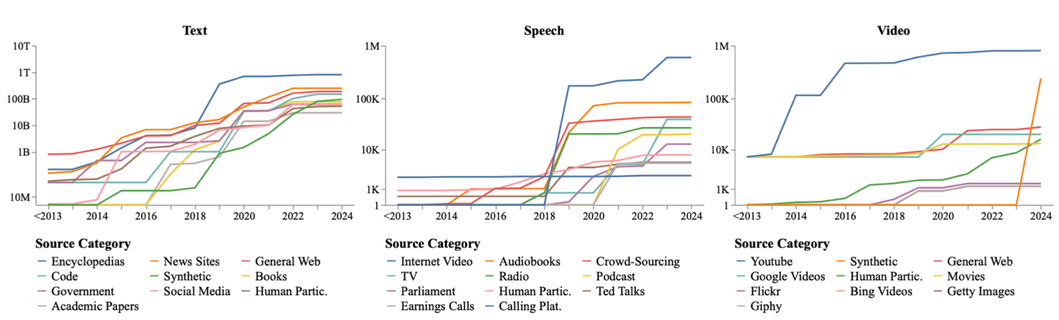
Summary:

- To understand the AI data ecosystem, we audit 3916 of the most popular, public ML datasets for text, speech, and video generation—summarized here
- We trace their **sources**, **license restrictions**, **languages**, and the **geographical origins** of the organizations that published them. (All open sourced!)
- We outline the three major observations on data sourcing, restrictions, and representation.

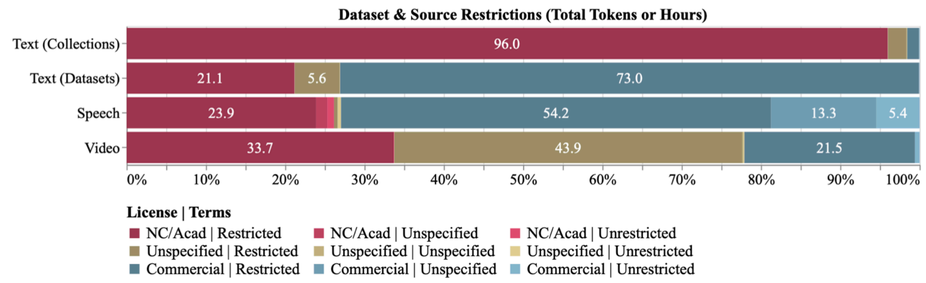
Our audit spans ~4k datasets, from 798 sources, 659 orgs, in 67 countries, & 608 languages.

	Datasets		Sources		Creator Orgs		Languages		Tasks	Licenses
	#	Size	#	Domains	#	Countries	#	Families		
Text	3717	2.1T	713	23	534	60	502	21	395	50
Speech	95	775k	51	16	124	29	260	36	18	19
Video	104	1.13M	44	24	101	23	-	-	33	11
Total	3916	-	798	83	659	67	608	37	443	55

1 Result: Multimodal data is increasingly sourced from the web, social media (YouTube), or synthetically generated

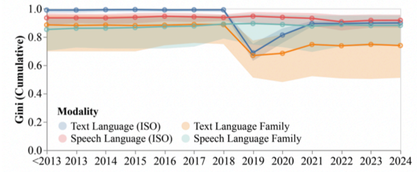


2 Result: 20-33% of datasets are non-commercially licensed, but >80% have undocumented source restrictions.

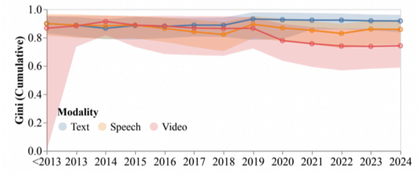


3 Result: Geographical and linguistic representation have not improved for a decade, across the data ecosystem.

	Africa	Asia	Europe	N America	Oceania	S America
By Count						
Text	0.3	13.4	24.0	61.5	0.7	0.2
Speech	3.6	35.7	30.4	30.4	0.0	0.0
Video	0.0	25.2	24.4	48.0	0.8	1.6
By Tokens or Hours						
Text	0.0	6.1	55.4	38.4	0.1	0.0
Speech	0.1	38.8	18.8	42.4	0.0	0.0
Video	0.0	23.1	22.0	38.2	16.7	0.1



(b) Gini-coefficient for Language Representation



(d) Gini-coefficient for Geographical Representation

While the amount of data from under-represented creators and languages increases each year, their relative representation remains consistently western-centric, with no significant improvements from >0.7 Gini coefficients.

Check out our paper here: