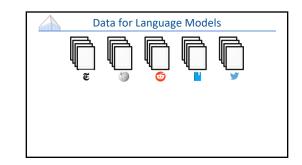
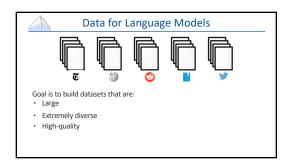


Building Datasets





The Web as a Data Source

Can we just scrape web data for training? Yes!

Ж

	The Web as a	Data Source
Can we j	ist scrape web data for training	^{3? Yes!} Common Crawl
Challeng	es: w to scrape effectively → C	
	ns of duplicate data Ded	
	s of junk, typos, nonsense 🛁	
		dit links / Wikipedia refs / use classifi
► No	n-english data 📥 Language	detector

Other Data Sources

- Code and Math
 GitHub, StackOverflow, ...
- Academic and Technical Works
 o arXiv, bioRxiv, PubMed, textbooks, ...
- Books
 Project Gutenberg, Libgen, ...
- Wikipedia

 \wedge

Examples of Datasets

WebText (GPT-2)

- GPT-2 dataset (not publicly released)
 o replicated in public as "OpenWebText"
- Take outbound links from Reddit posts with 3+ karma



- Various filtering stages (lang detect, code removal, bad words)
- Used to train T5 and other models
- 806gb of text (~156 billion words)

Th Th	e Pile	
Component	Raw Size	Weight
Pile-CC	227.12 GiB	18.11%
PubMed Central	90.27 GiB	14.40%
Books3 [†]	100.96 GiB	12.07%
OpenWebText2	62.77 GiB	10.01%
ArXiv	56.21 GiB	8.96%
Github	95.16 GiB	7.59%
FreeLaw	51.15 GiB	6.12%
Stack Exchange	32.20 GiB	5.13%
USPTO Backgros	nds 22.90 GiB	3.65%
PubMed Abstract	19.26 GiB	3.07%
Gutenberg (PG-1	0 [†] 10.88 GiB	2.17%
OpenSubtitles [†]	12.98 GiB	1.55%
Wikipedia (en) [†]	6.38 GiB	1.53%
DM Mathematics	7.75 GiB	1.24%
Ubantu IRC	5.52 GiB	0.88%
BookCorpus2	6.30 GiB	0.75%
EuroParl [†]	4.59 GiB	0.73%
HackerNews	3.90 GiB	0.62%
YoutubeSubtitles	3.73 GiB	0.60%
PhilPapers	2.38 GiB	0.38%
NIH ExPorter	1.89 GiB	0.30%
Enron Emails [†]	0.88 GiB	0.14%
The Pile	825.18 GiB	

Risks, Harms, and Biases

Representation and Fairness

Data defines model behavior

- Today's LM datasets are biased in various ways:
- Anti muslim
 Reddit datasets are liberal leaning
- Stereotypical gender roles
- o

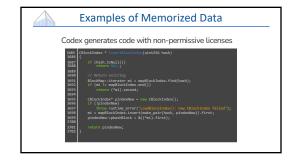
Models inherit these biases

Toxicity and Safety

- Today's LM datasets are full of toxic junk:
- Hate speech
- Racist posts
- Toxic comments
 Misinformation / fake news
- Models inherit these aspects













Ongoing Lawsuits Regarding Copyright Data

We're investigating a potential lawsuit against GitHub Copilot for violating its legal duties to open-source authors and end

Ongoing Lawsuits Regarding Copyright Data

We're investigating a potential lawsuit against GitHub Copilot for violating its legal duties to

Open-SOUICE Getty Images is suing the creators of Al art tool Stable Diffusion for scraping its content

Ongoing Lawsuits Regarding Copyright Data

We're investigating a potential lawsuit against GitHub Copilot for violating its legal duties to open-source**Getty mages is suing the creat**

Open-sourceGettv Images is suing the creators of AI We've filed a lawsuit challenging raping its Stable Diffusion, a 21st-century collage tool that violates the rights of artists.

