

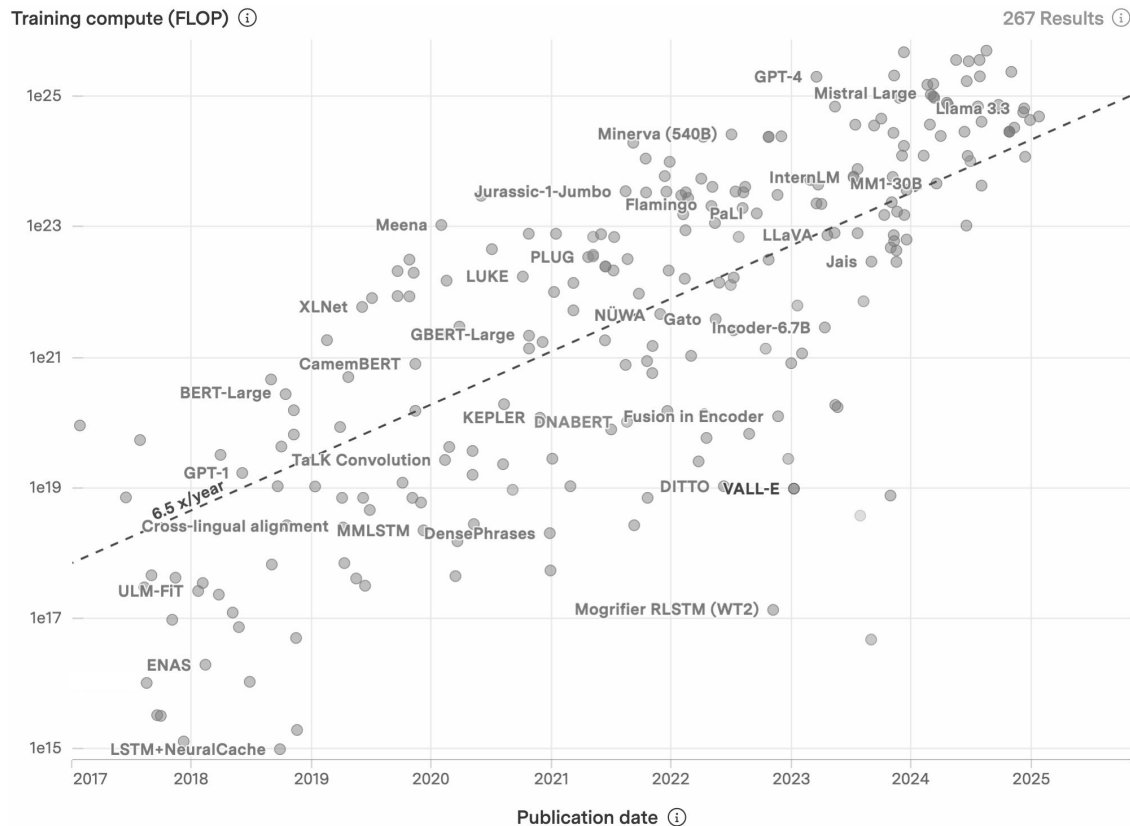
The most expensive
part of an LLM *should*
be its training data

Colin Raffel

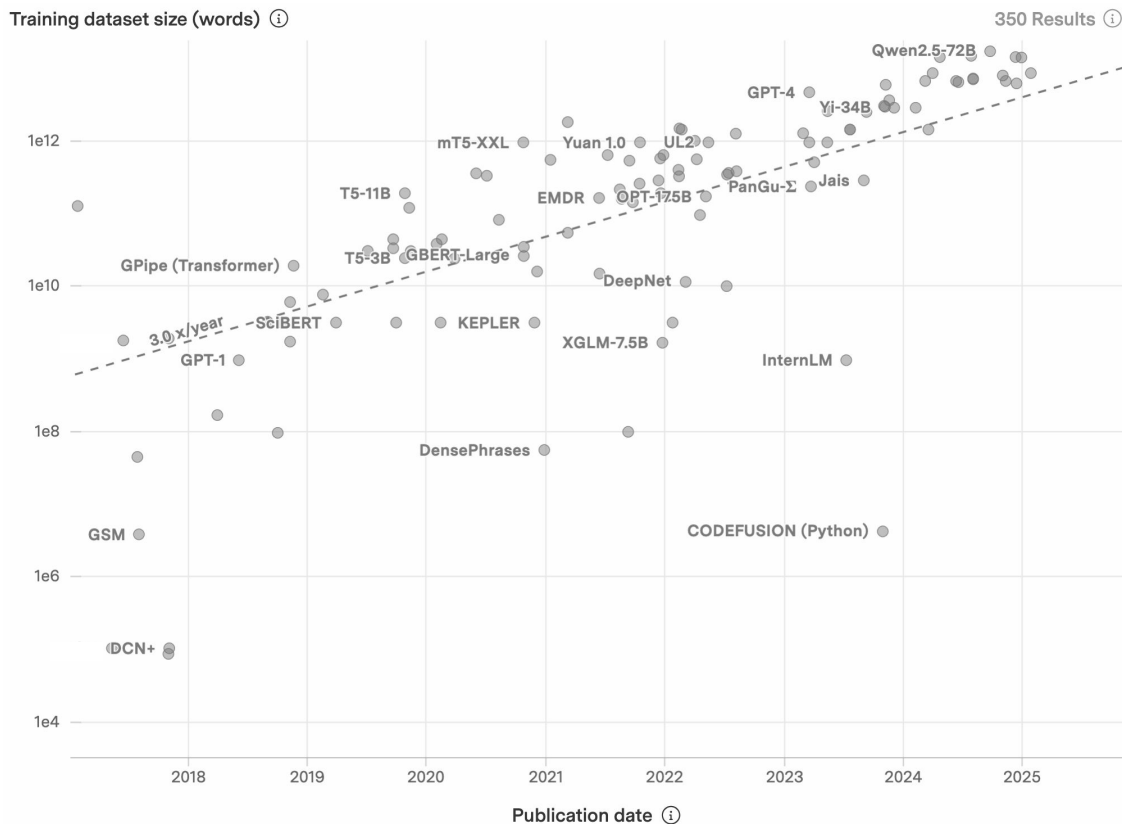
(presenting work primarily by Nikhil Kandpal)



Language model training costs have increased ~6.5x each year

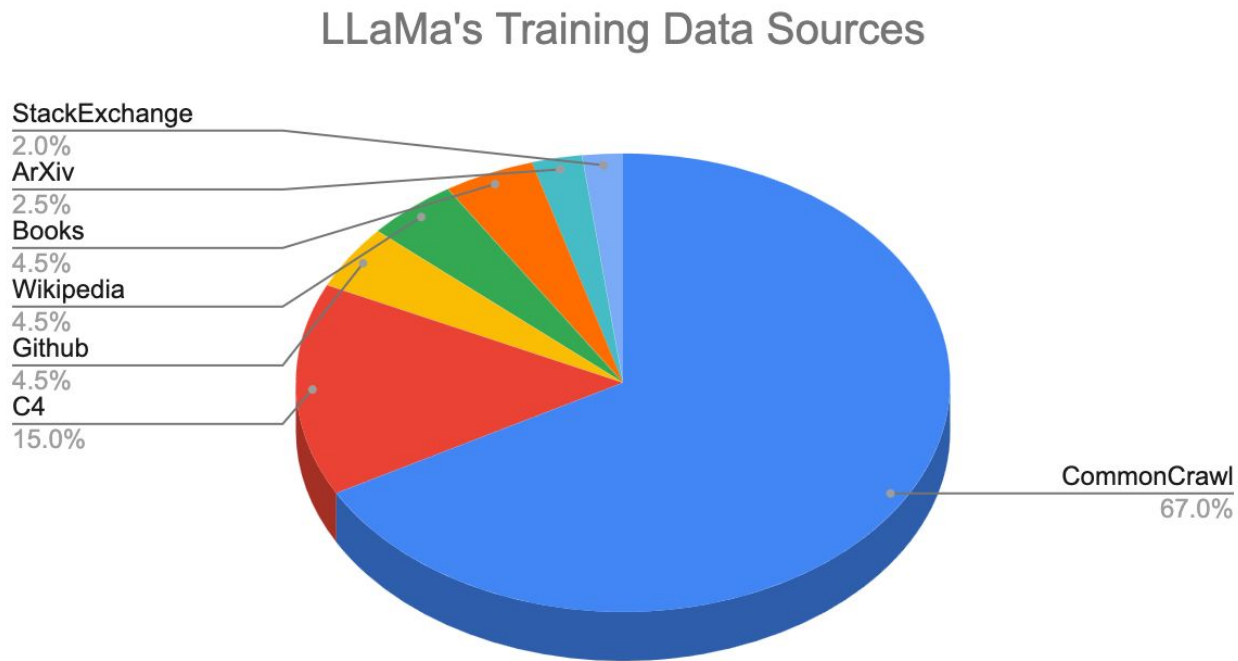


Increasing model sizes leads to increasing training dataset sizes



From "Data on Notable AI Models" by EpochAI

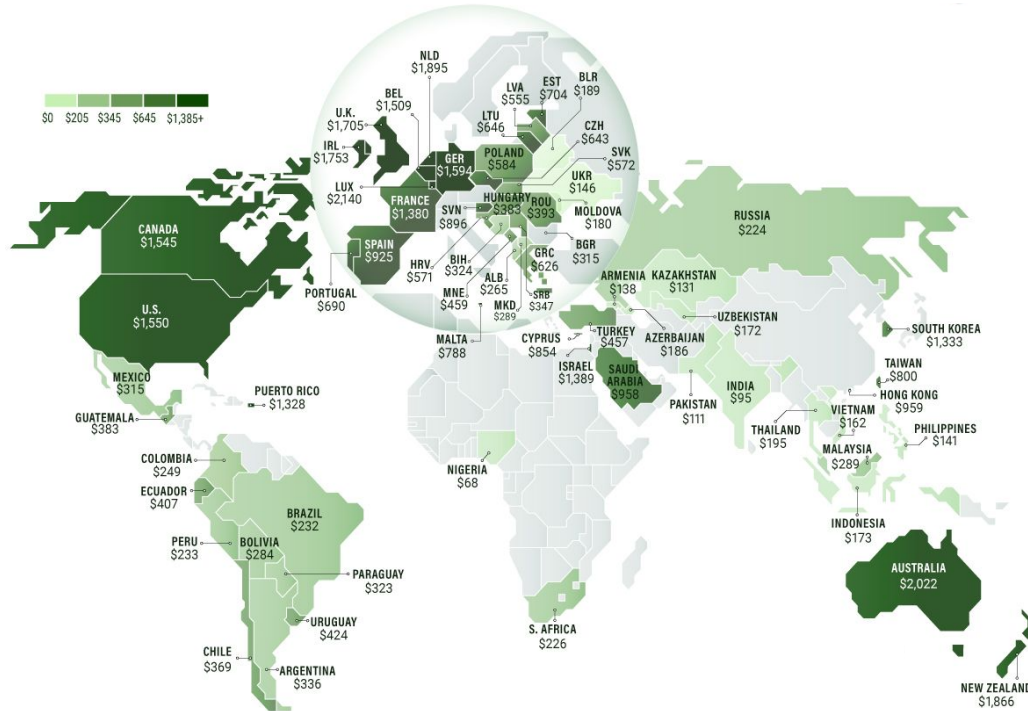
Typical training datasets comprise many sources



Writing text involves human labor

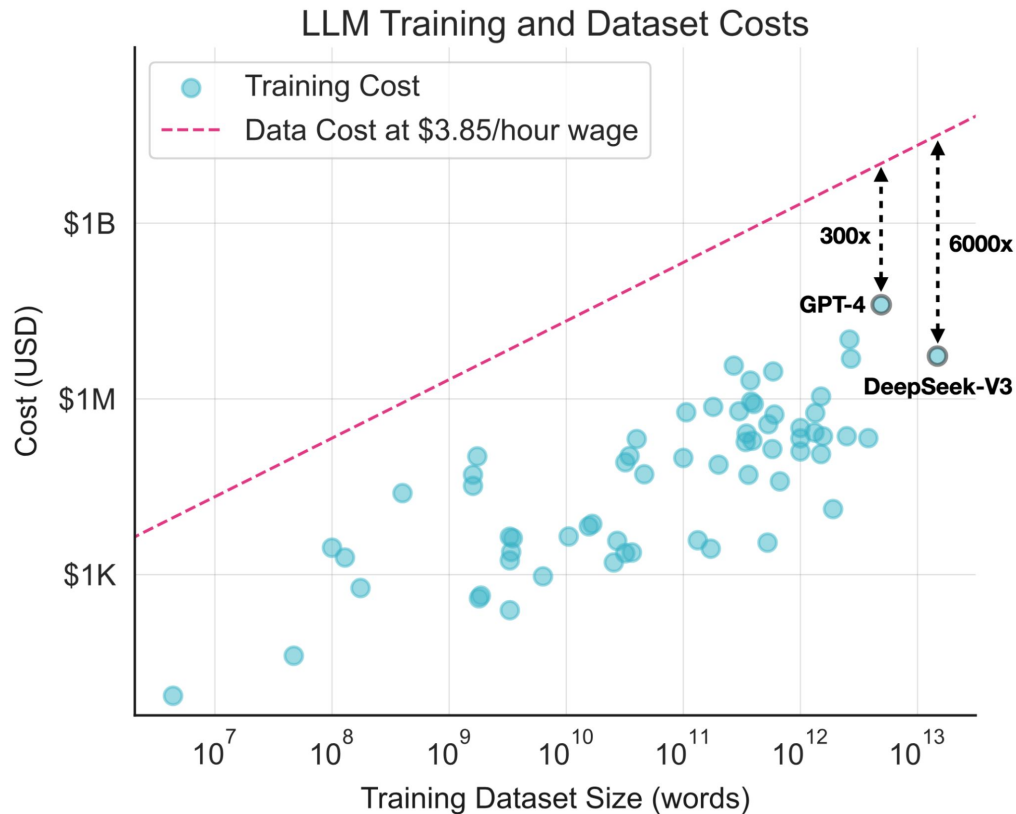
Writing Medium	Typical Length (words)	Estimated Writing Time
Blog Post	2,000	1.1 hours
Academic Paper	5,000	2.8 hours
Novel	70,000	1.6 days
Textbook	300,000	1 week
Encyclopedia Britannica	40,000,000	2.5 years

Human labor has a cost



From <https://www.visualcapitalist.com/minimum-wage-around-the-world/>

Training data would cost much more than training if annotators were paid



From "The Most Expensive Part of an LLM should be its Training Data" by Kandpal and Raffel

Copyright-related lawsuits against OpenAI/Microsoft as of March 2024

Coders

1. Joseph Saveri Firm: [overview](#), [complaint](#)

Writers

2. Joseph Saveri Firm: [overview](#), [complaint](#)
3. Authors Guild & Alter: [overview](#), [complaint](#)
4. Nicholas Gage: [overview & complaint](#)

Media

5. New York Times: [overview](#), [complaint](#)
6. Intercept Media: [overview](#), [complaint](#)
7. Raw Story & Alternet: [overview](#), [complaint](#)
8. Denver Post & seven others: [overview](#), [complaint](#)
9. Center for Investigative Reporting: [overview](#), [complaint](#)

Idea 1:

Train only on "permissively
licensed" text

“... it would be impossible to train today’s leading AI models without using copyrighted materials.”

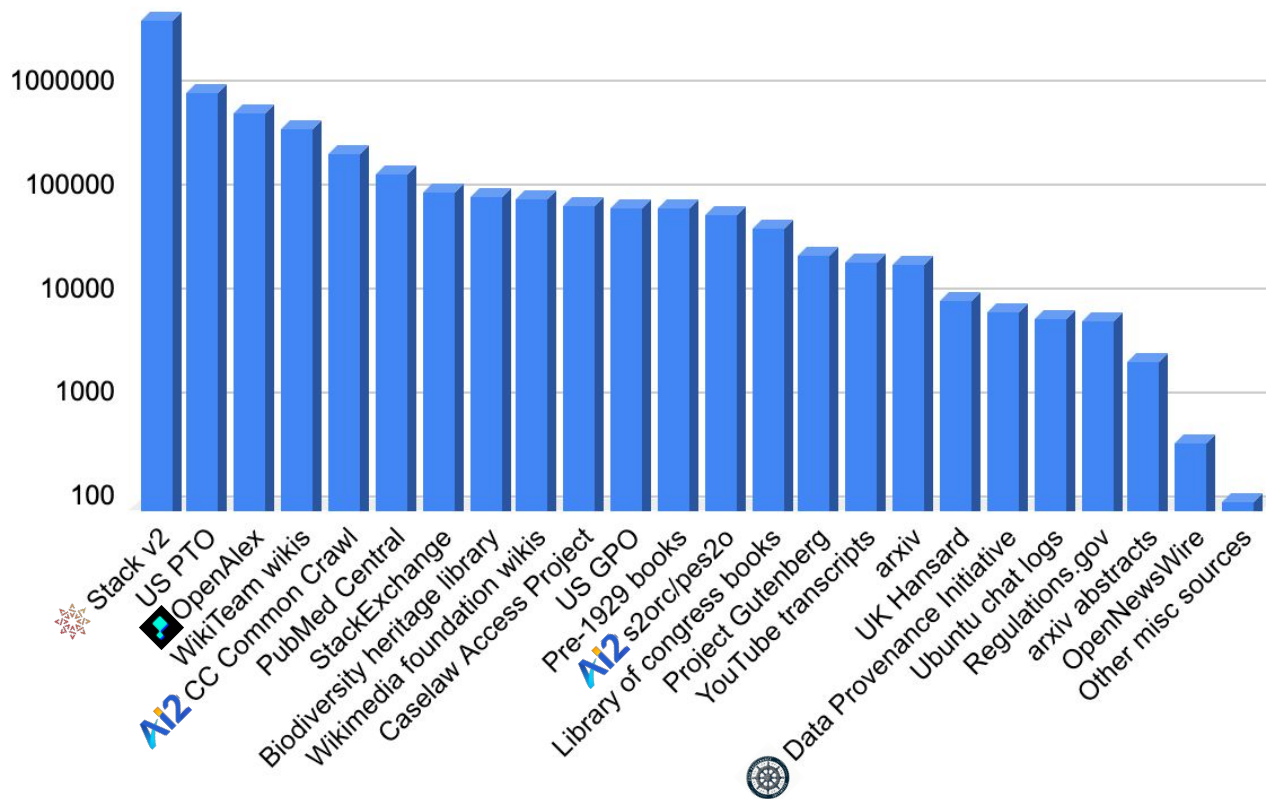
– [OpenAI](#)

... let's try!

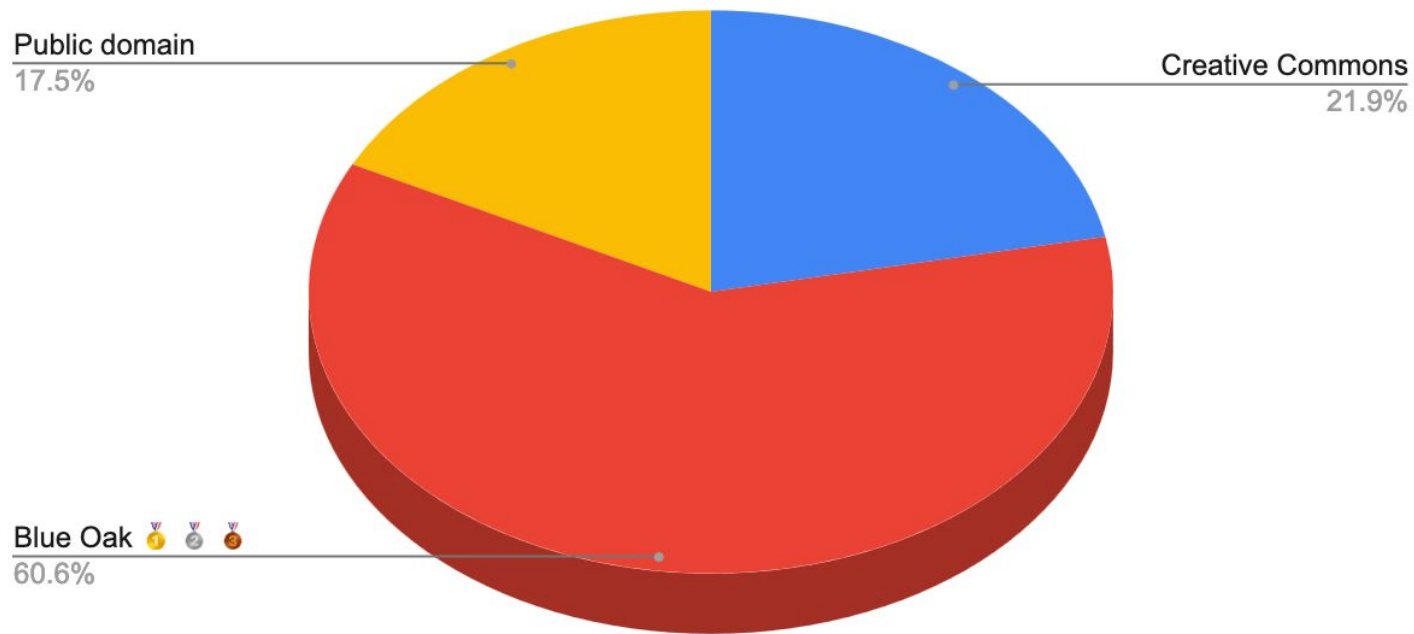
What is "permissively licensed" (to us)?

- I'm not a lawyer, but I know people who know lawyers
- Public domain/CC0
 - Mostly very old and/or governmental text
- Creative Commons-Attribution (CC-BY, CC-BY-SA)
 - Non-commercial (NC) considered non-permissive
 - "No derivative works" is ambiguous
 - (so is attribution, sort of...)
- Blue Oak Council gold/silver/bronze licenses
 - BSD, MIT, Apache, etc.
- Licenses "equivalent" to the above
- https://github.com/r-three/common-pile/blob/main/licensed_pile/licenses.py

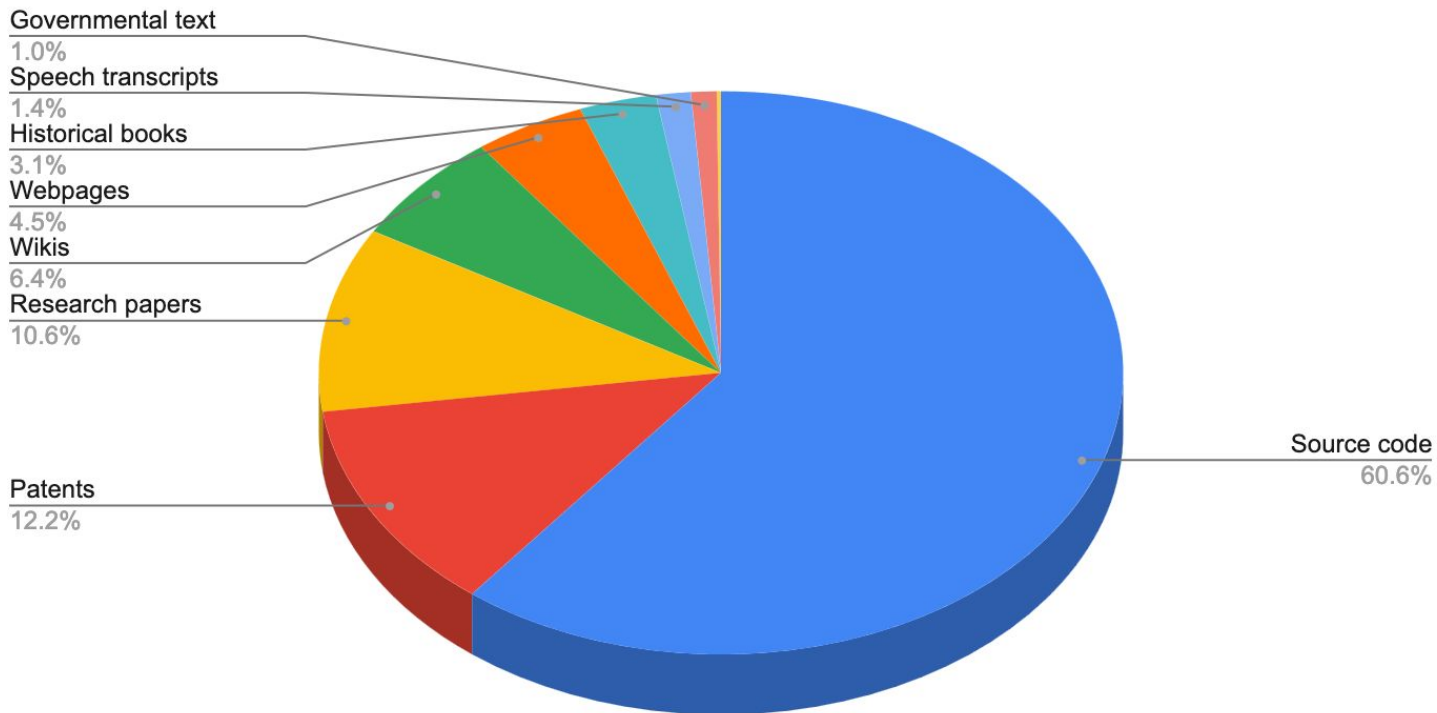
Common Pile raw source sizes (in UTF-8 bytes)



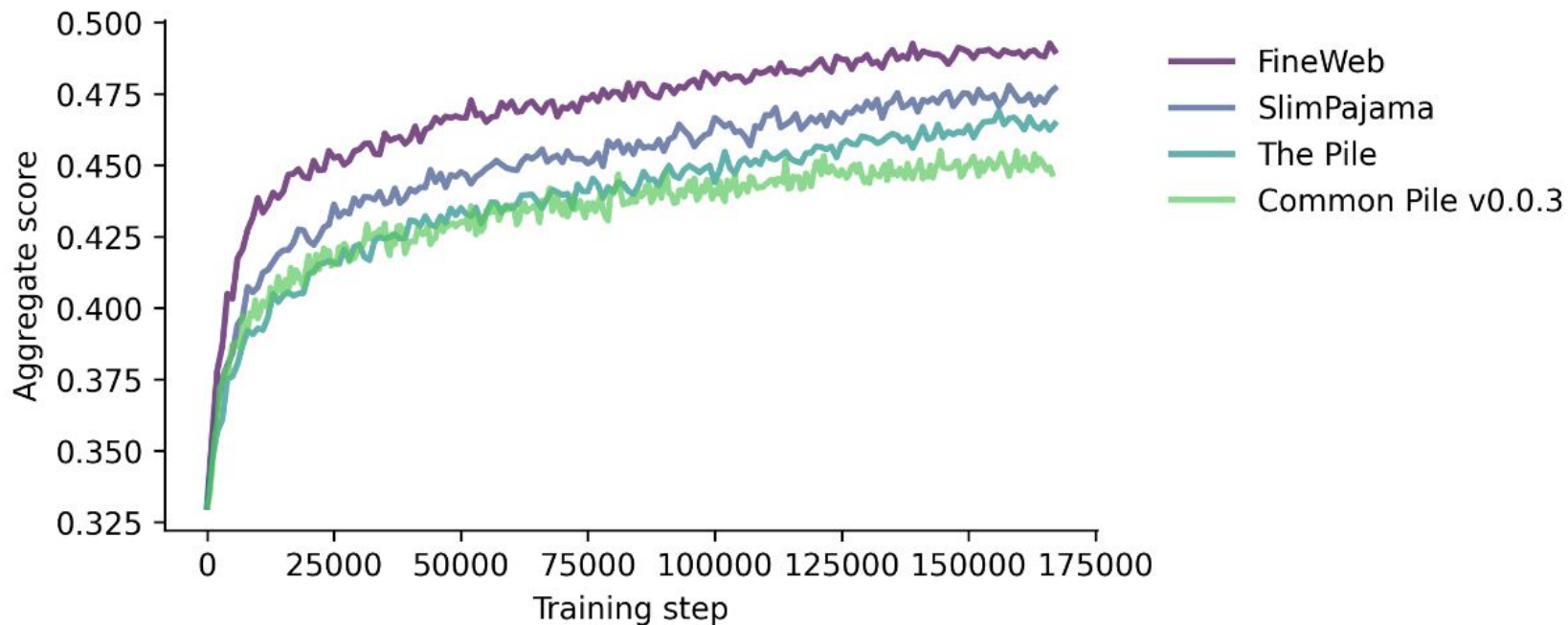
Proportion of licence types



Proportion of source types



How it's going



Try it out!

 **Hugging Face**

 Models  Datasets  Spaces  Posts  Docs  Enterprise Pricing  

Datasets:  nkandpa2 / **common-pile-filtered**   like 0

Modalities:  Text Formats:  json Size: 1B - 10B Libraries:  Datasets  Dask  Croissant

 **Dataset card**  Viewer  Files and versions  Community 2

 This dataset has 1 file scanned as unsafe. [Show files](#)

Dataset Viewer (First 5GB)   Auto-converted to Parquet  API  Embed  Full Screen Viewer

Subset (26)
default · ~1.18B rows (showing the first 2.53M) 

Split (1)
train · ~1.18B rows (showing the first 2.53M) 

 SQL  Console

added string · lengths	created timestamp[s]	id string · lengths	metadata dict	source string · classes
				
19 26		9 16		2 values
2024-08-12T18:16:26.585370	2007-04-02T19:18:42	0704.0001	{ "authors": "C. Bal\ 'azs, E. L. Berger, P...	arxiv-abstracts
2024-08-12T18:16:26.585578	2007-03-31T02:26:18	0704.0002	{ "authors": "Ileana Streinu and Louis...	arxiv-abstracts
2024-08-12T18:16:26.585644	2007-04-01T20:46:54	0704.0003	{ "authors": "Hongjun Pan", ...	arxiv-abstracts
2024-08-12T18:16:26.585690	2007-03-31T03:16:14	0704.0004	{ "authors": "David Callan", ...	arxiv-abstracts
2024-08-12T18:16:26.585725	2007-04-02T18:09:58	0704.0005	{ "authors": "Wael Abu-Shammala and Alberto...	arxiv-abstracts
2024-08-12T18:16:26.585760	2007-03-31T04:24:59	0704.0006	{ "authors": "Y. H. Pong and C. K. Law", ...	arxiv-abstracts

< Previous 1 2 3 ... 25,345 Next >

Downloads last month 1,150

 Use this dataset  Edit dataset card 

Size of the auto-converted Parquet files (First 5GB per split):
48.6 GB

Number of rows (First 5GB per split):
17,386,843

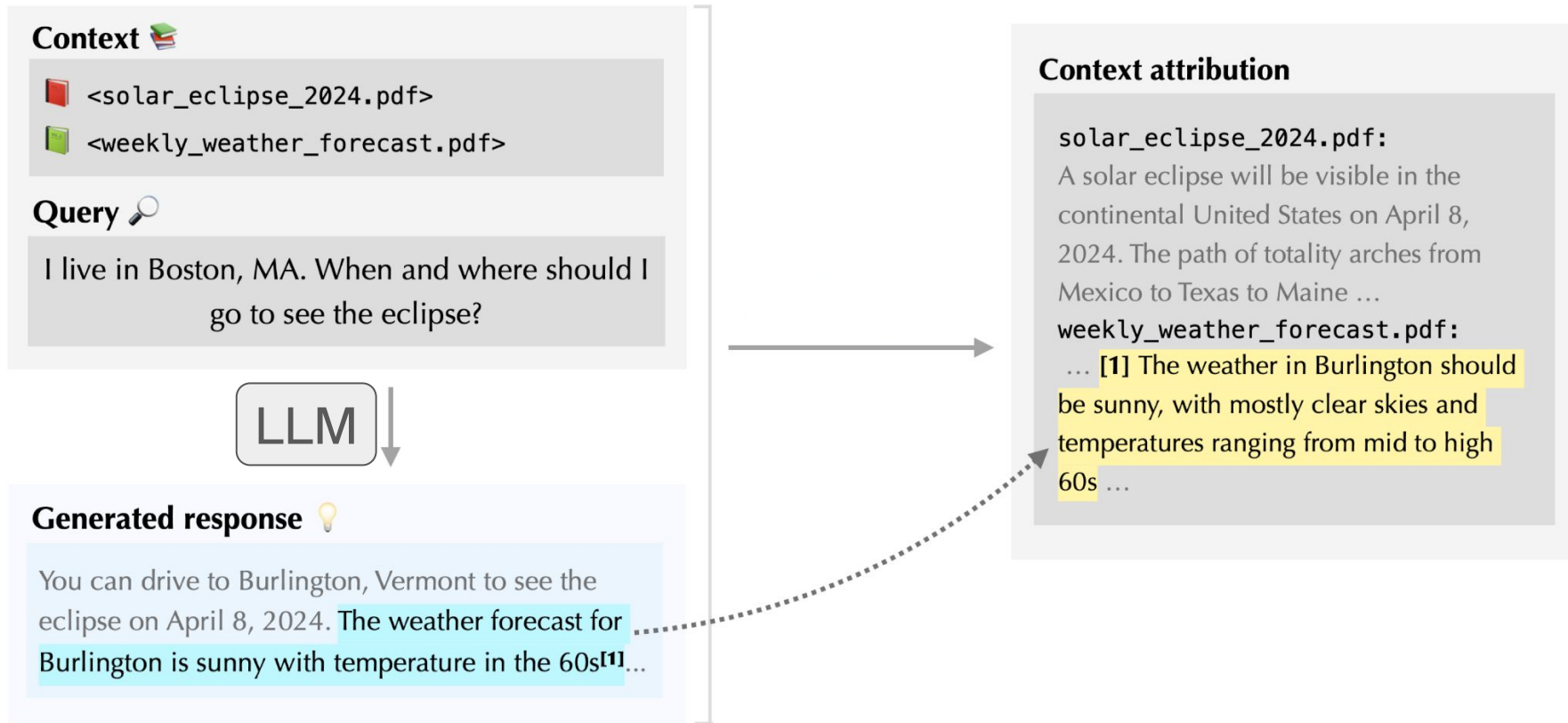
Estimated number of rows:
1,365,167,755

<https://huggingface.co/datasets/nkandpa2/common-pile-filtered>

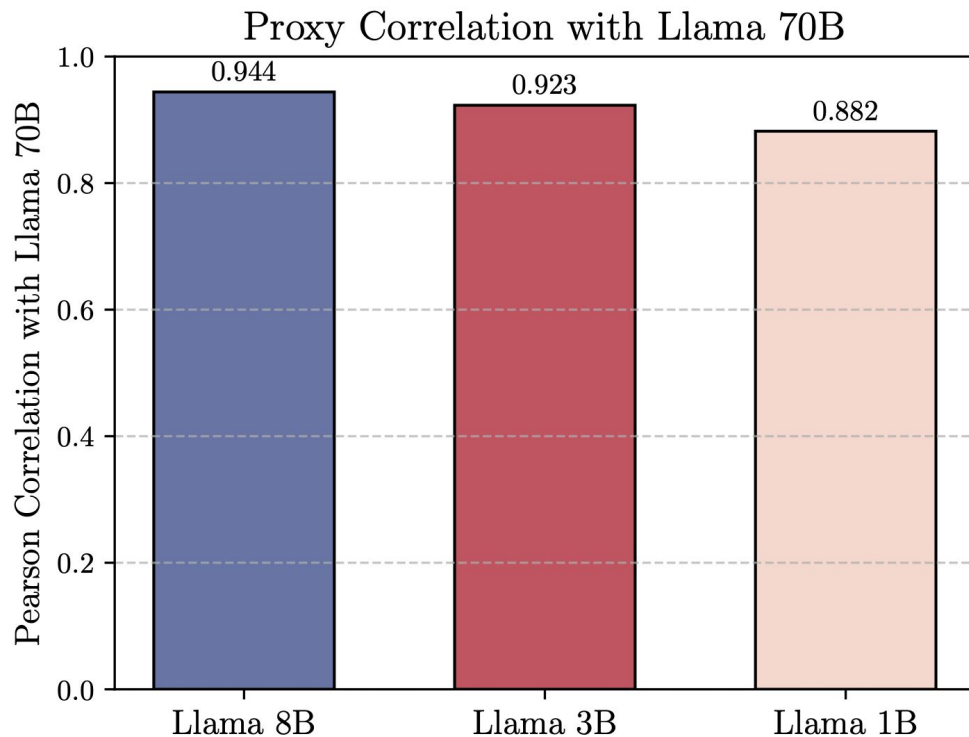
Idea 2:

Attribute predictions back
to training data

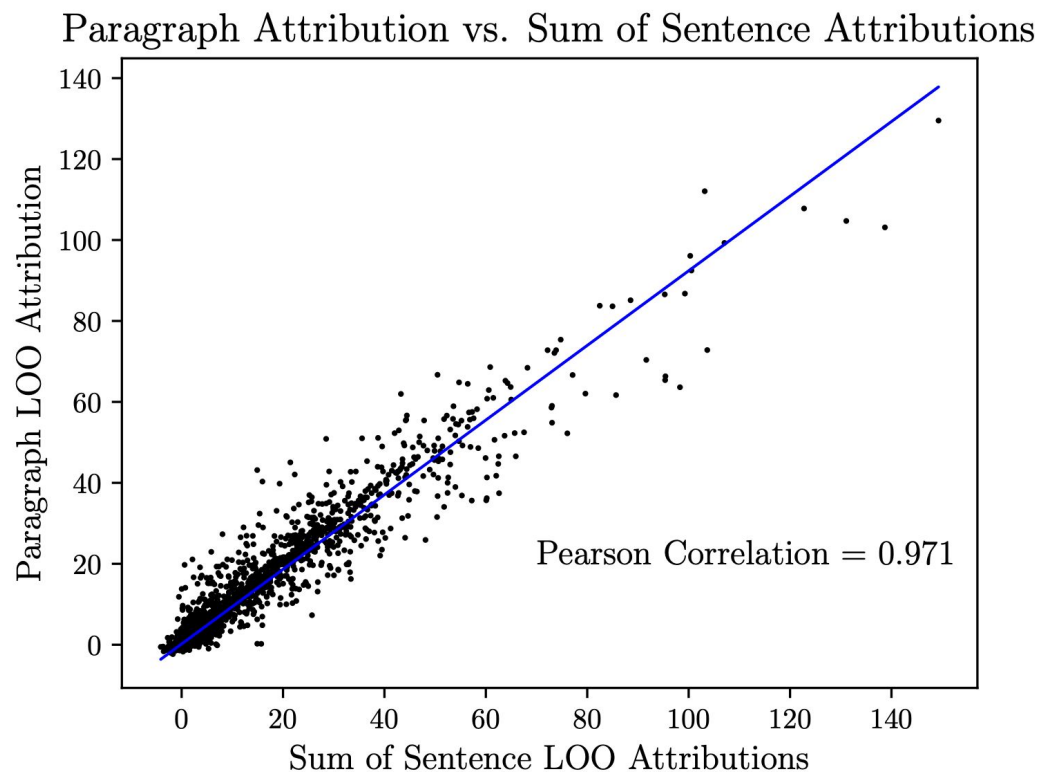
Context attribution finds influential spans in a model's input



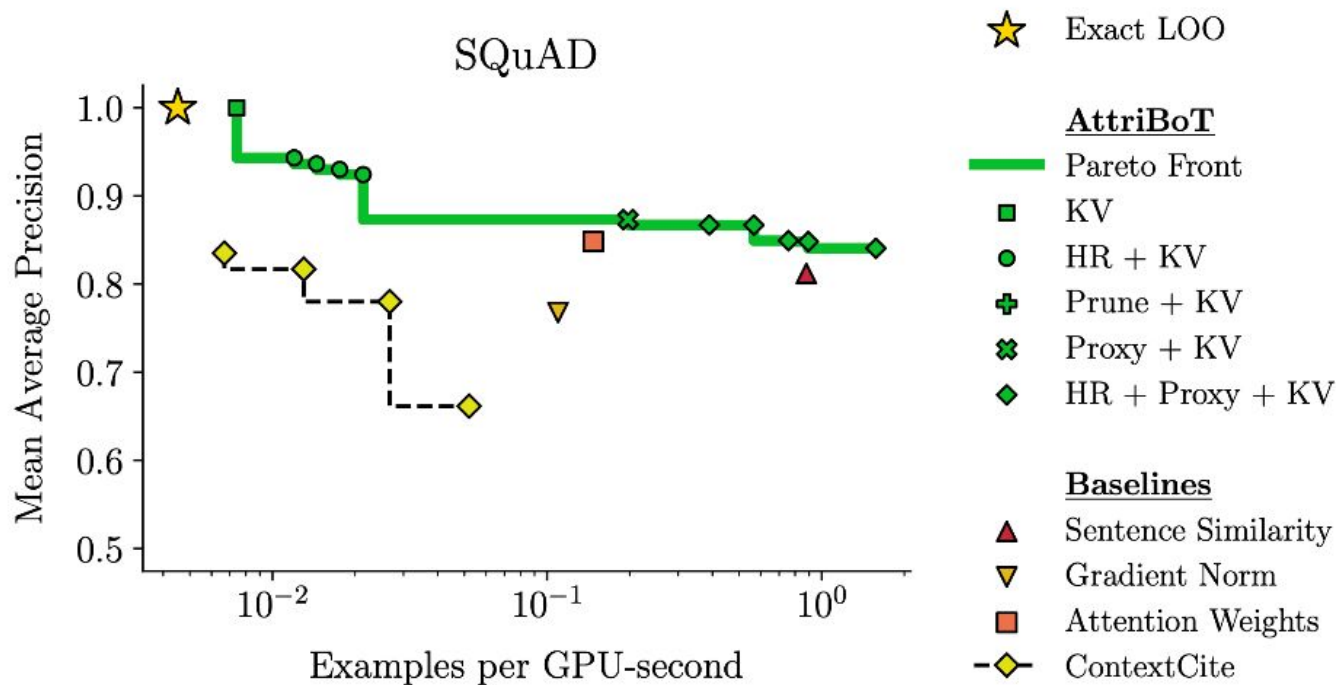
Observation 1: Smaller models provide similar attributions



Observation 2: Attributions combine linearly



AttriBoT is pareto-optimal for context attribution



Try it out!



AttriBoT

Public



Edit Pins



Watch

0



Fork

0



Star

5



main



Go to file



<> Code



FY-Liu Fixed gradnorm device issue.

e3ffcea · 3 months ago



context_attribution

Fixed gradnorm device issue.

3 months ago



context_attribution_simple.e...

Initial upload.

3 months ago



example

Initial upload.

3 months ago



README.md

Initial upload.

3 months ago



requirements.txt

Initial upload.

3 months ago



setup.py

Initial upload.

3 months ago

About



AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution

📖 Readme

📈 Activity

📋 Custom properties

☆ 5 stars

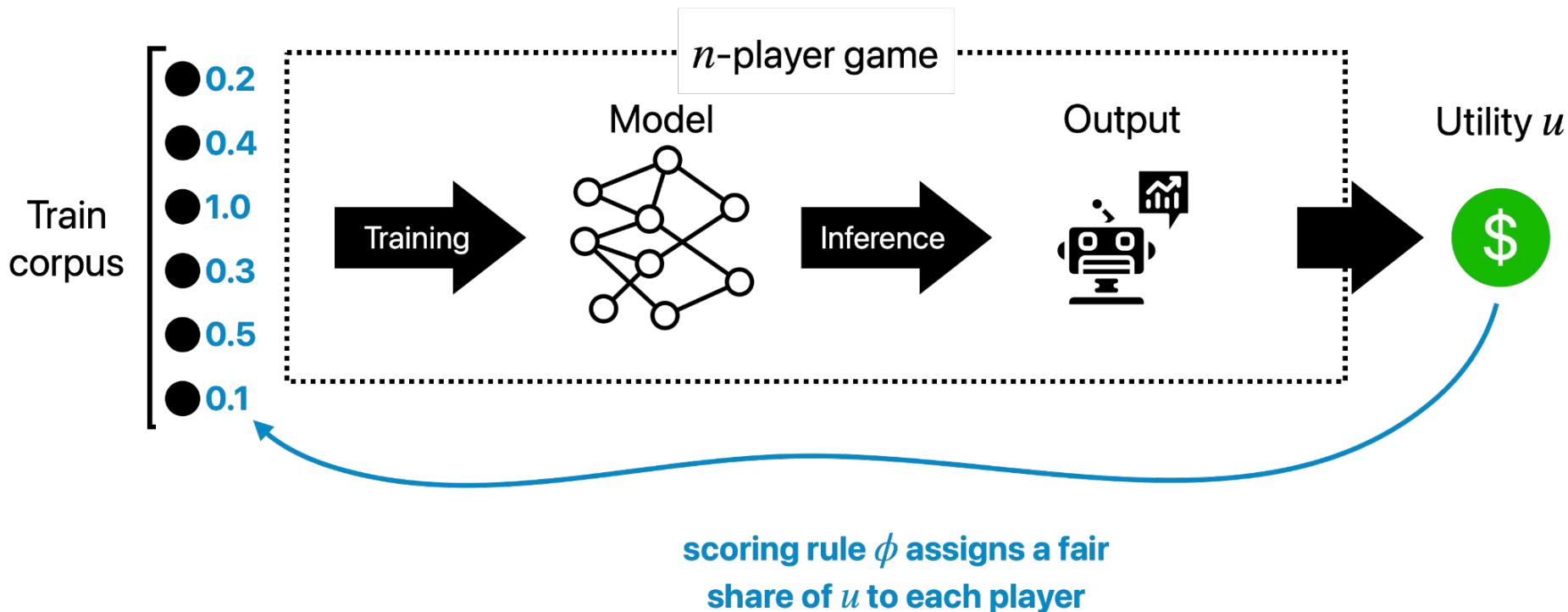
👁 0 watching

🍴 0 forks

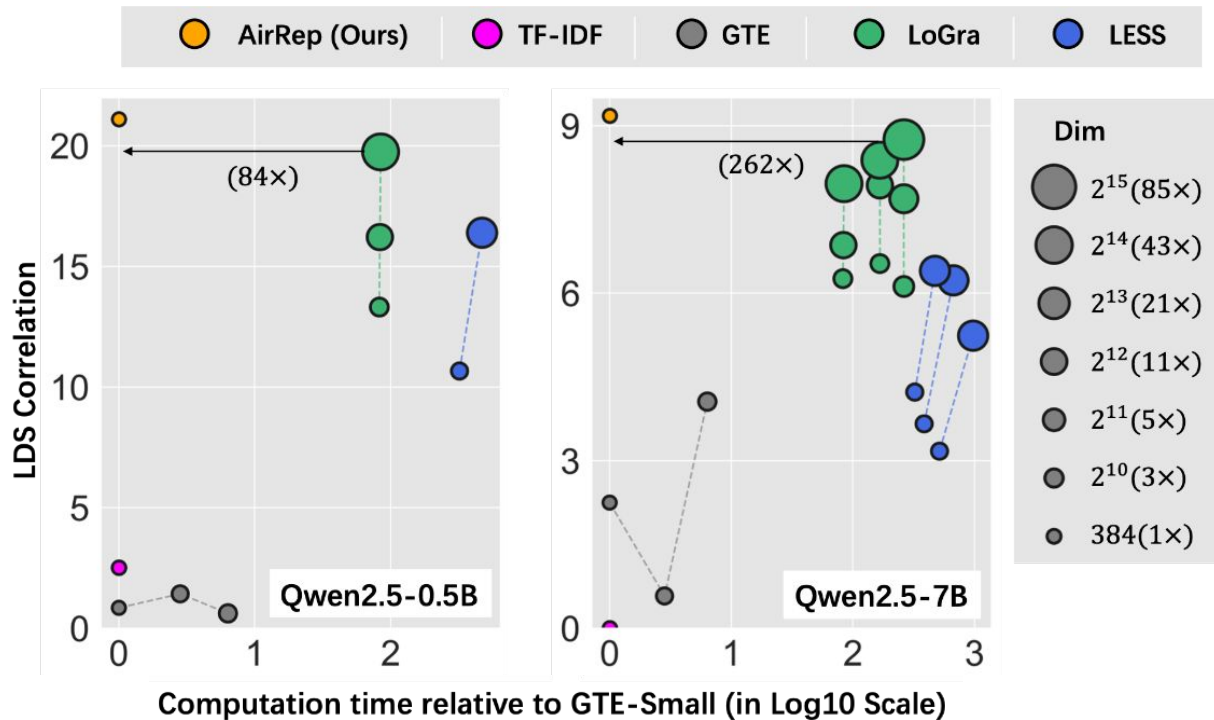
Report repository

From <https://github.com/r-three/AttriBoT>

Measuring training data influence



AirRep is pareto-optimal for group data influence



Thanks.

Please give me feedback:

<http://bit.ly/colin-talk-feedback>

craffel@gmail.com