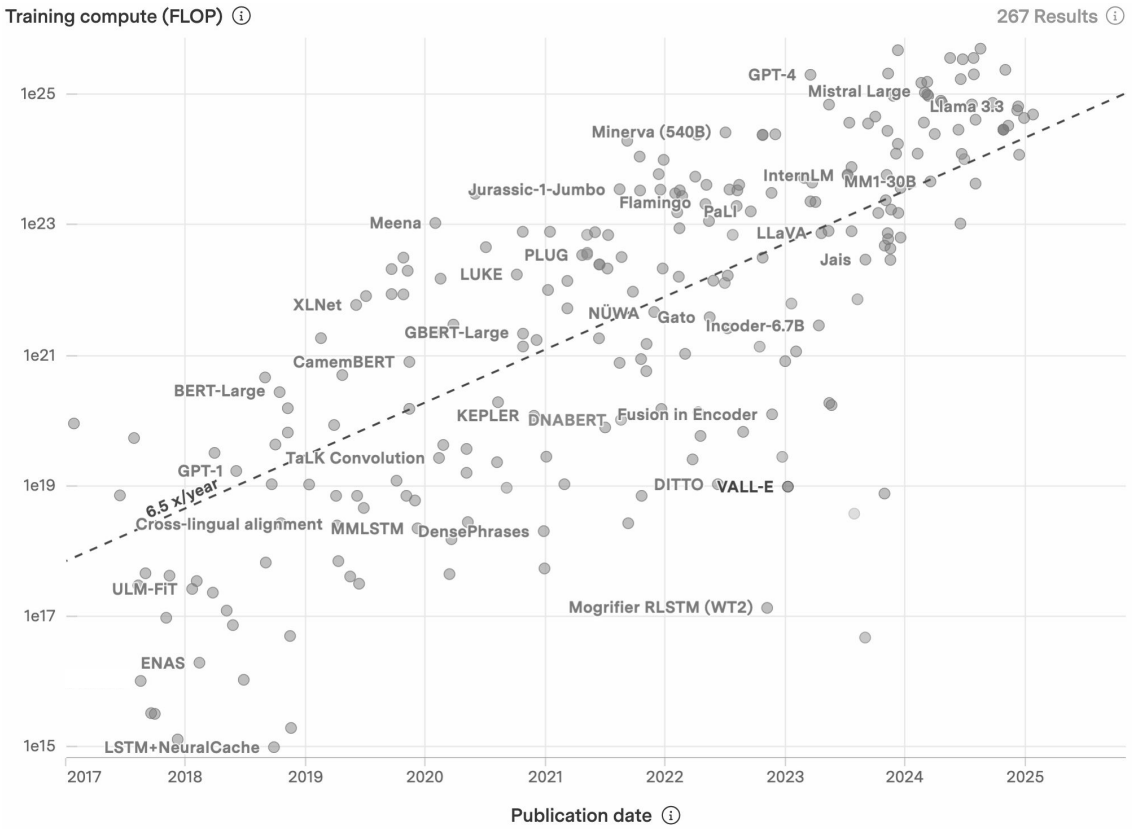
The most expensive part of an LLM *should* be its training data

Colin Raffel

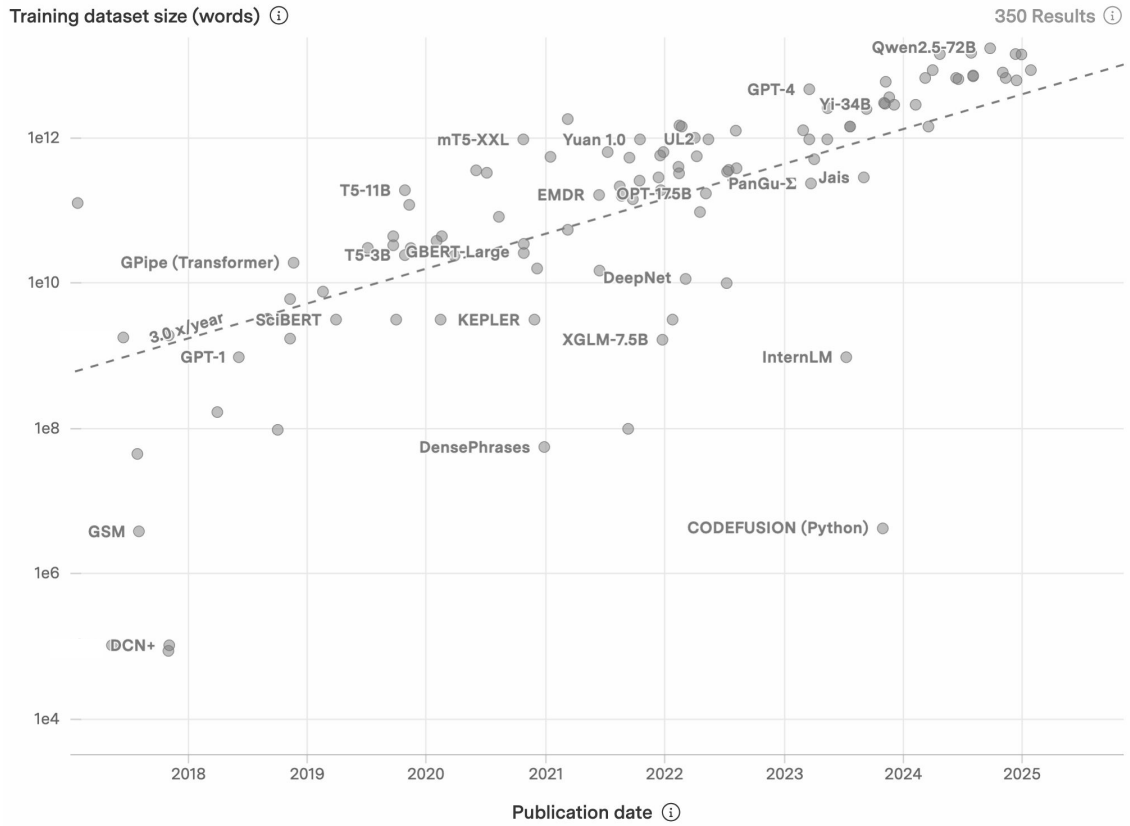
(presenting work primarily by Nikhil Kandpal)

Language model training costs have increased ~6.5x each year



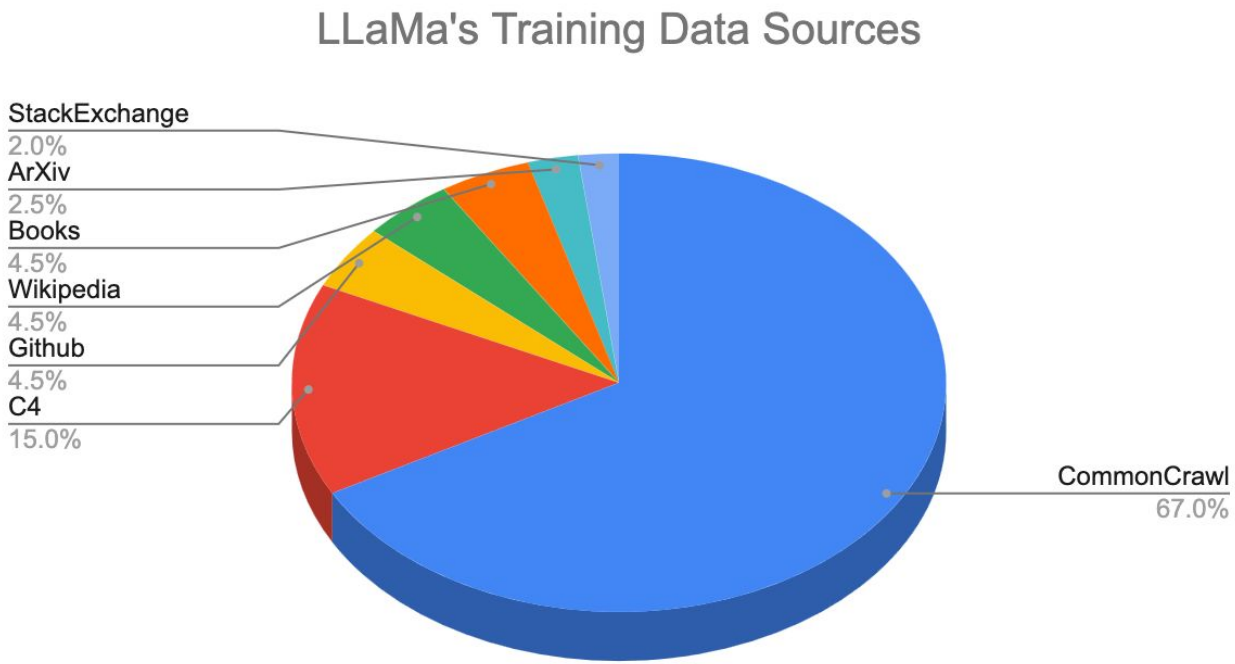
From "Data on Notable AI Models" by EpochAI

Increasing model sizes leads to increasing training dataset sizes



From "Data on Notable AI Models" by EpochAI

Typical training datasets comprise many sources



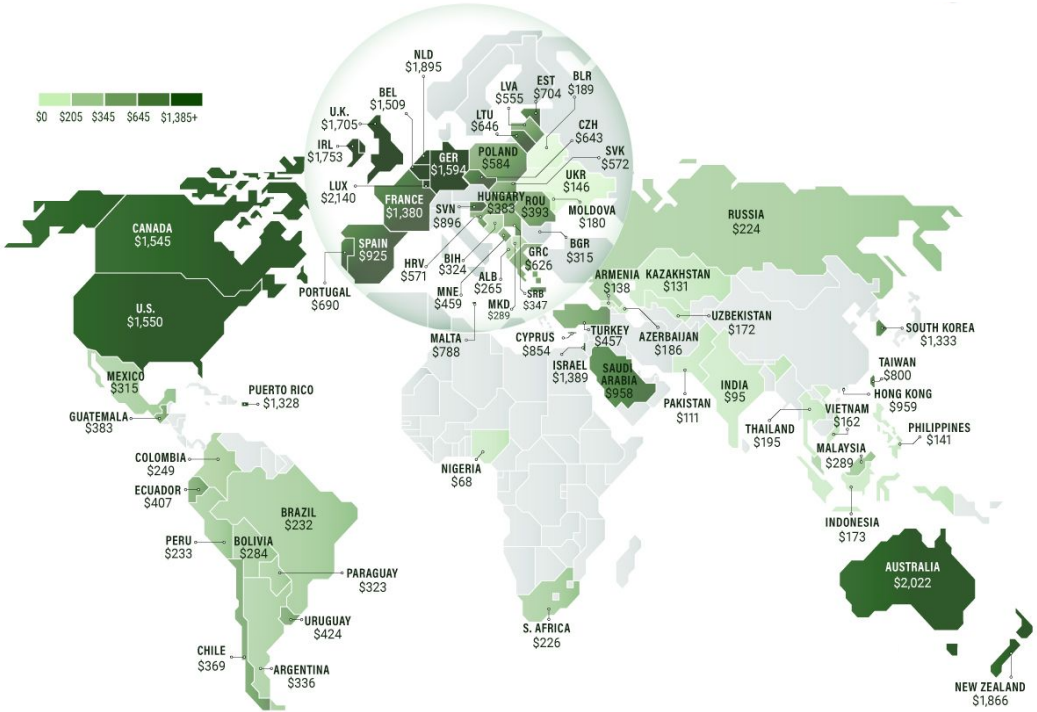
From "LLaMA: Open and Efficient Foundation Language Models" by Touvron et al.

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

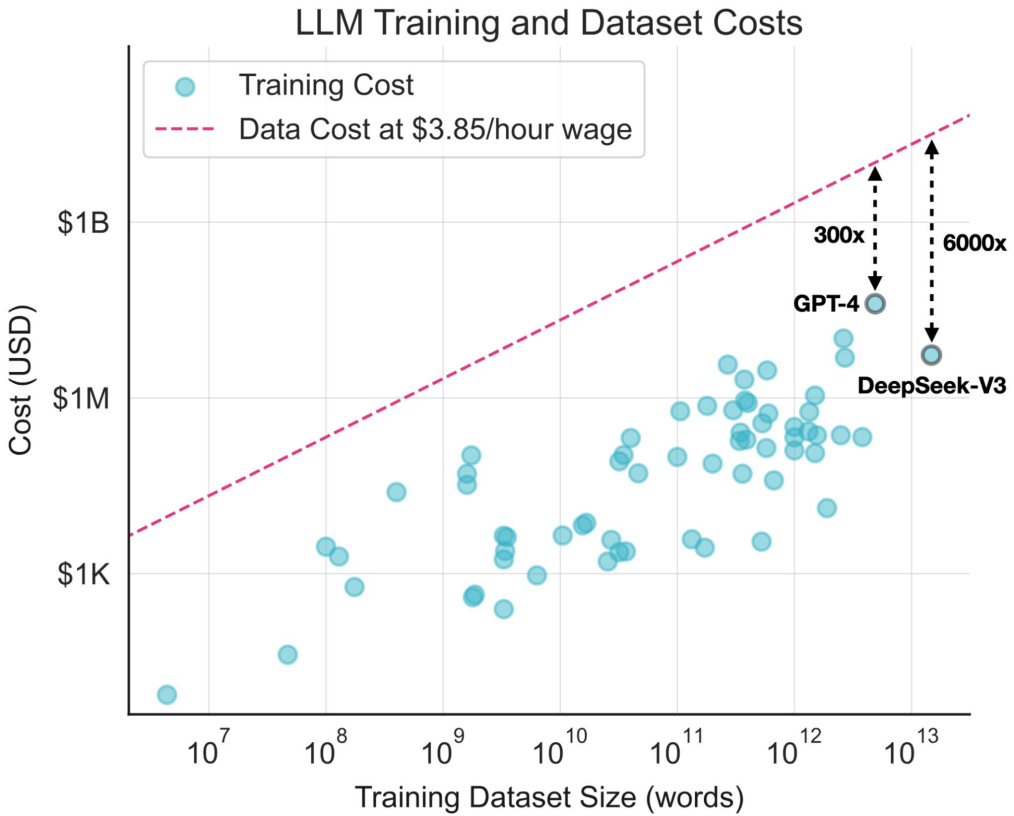
From "The Most Expensive Part of an LLM should be its Training Data" by Kandpal and Raffel

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

From https://forum.effectivealtruism.org/posts/EdBkBBFkCaHrw2iwL/twelve-lawsuits-against-openai

Idea 1:

Train only on "permissively licensed" text

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

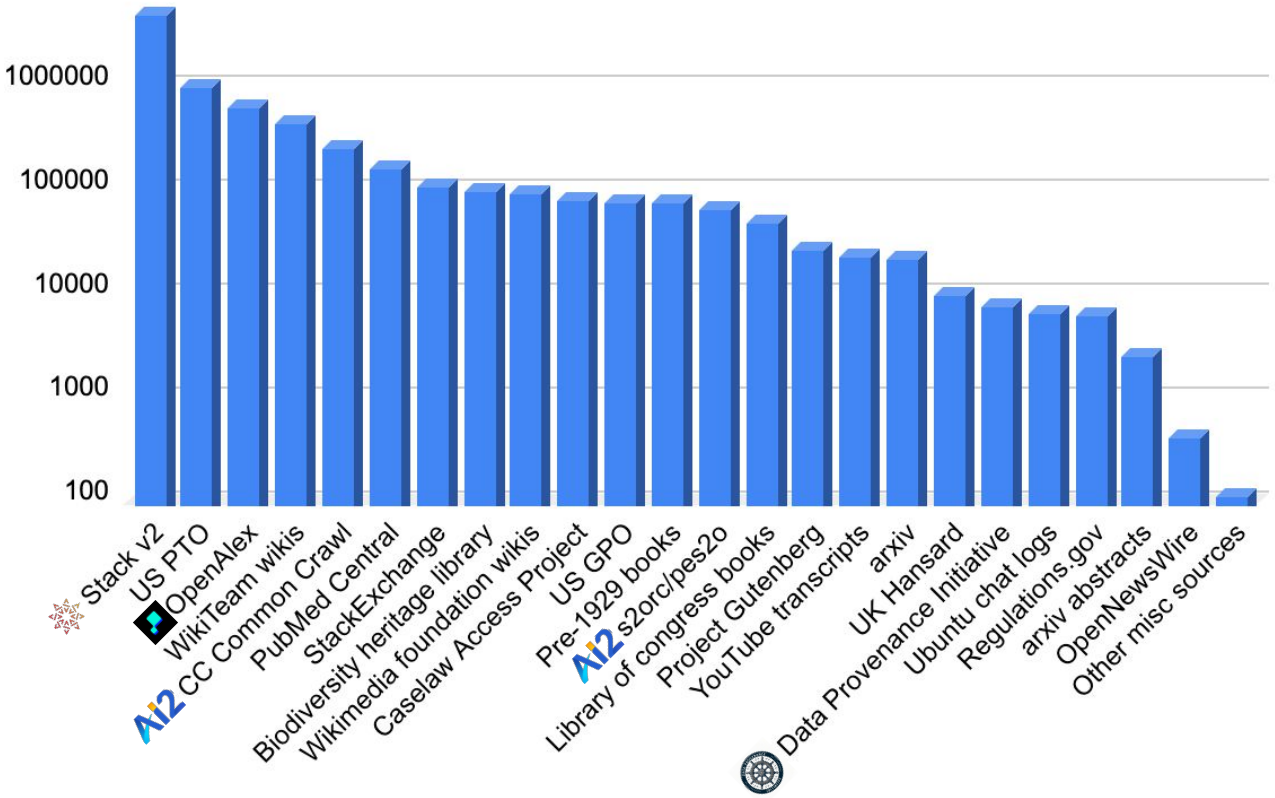
– OpenAI

… let's try!

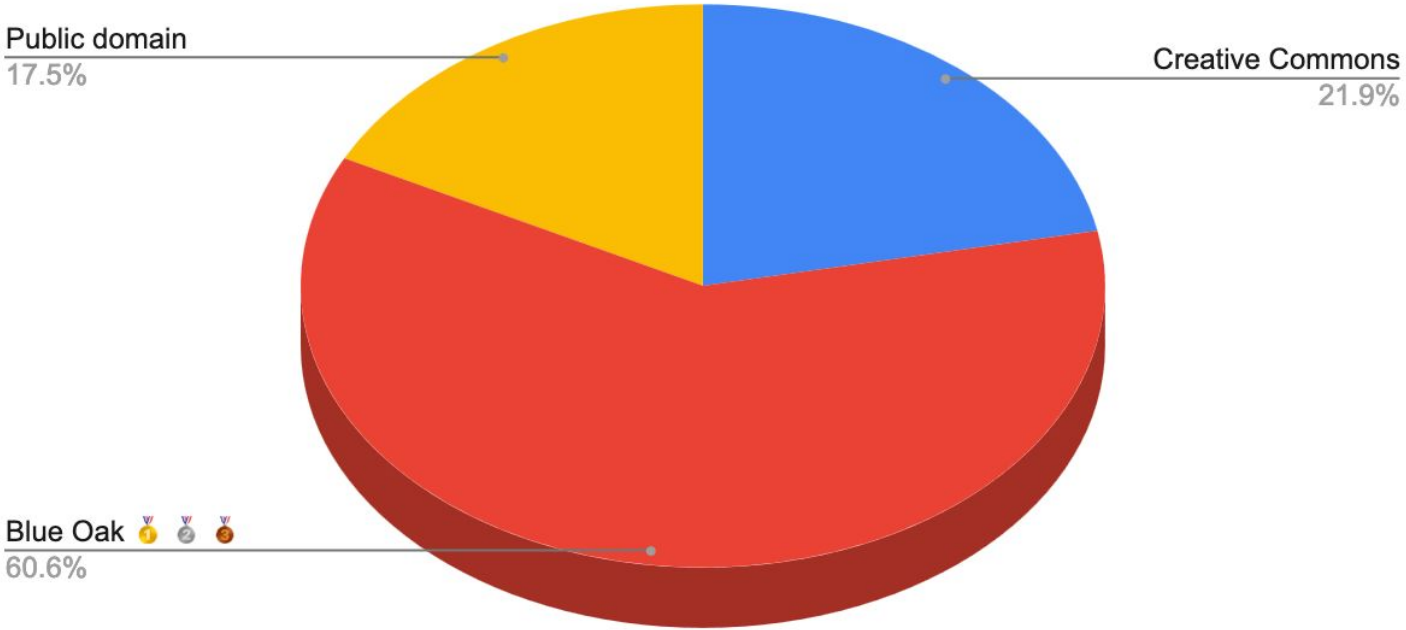
What is "permissively licensed" (to us)?

- 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

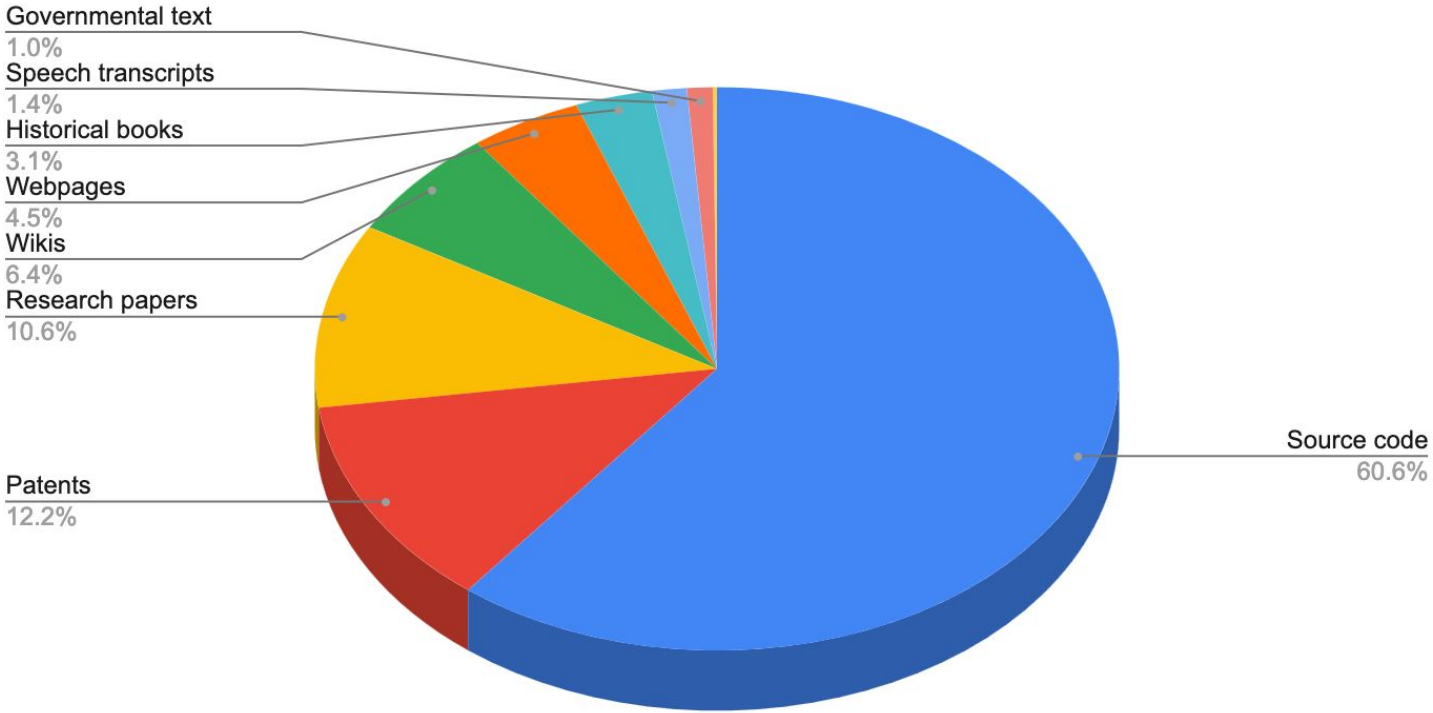
Common Pile raw source sizes (in UTF-8 bytes)



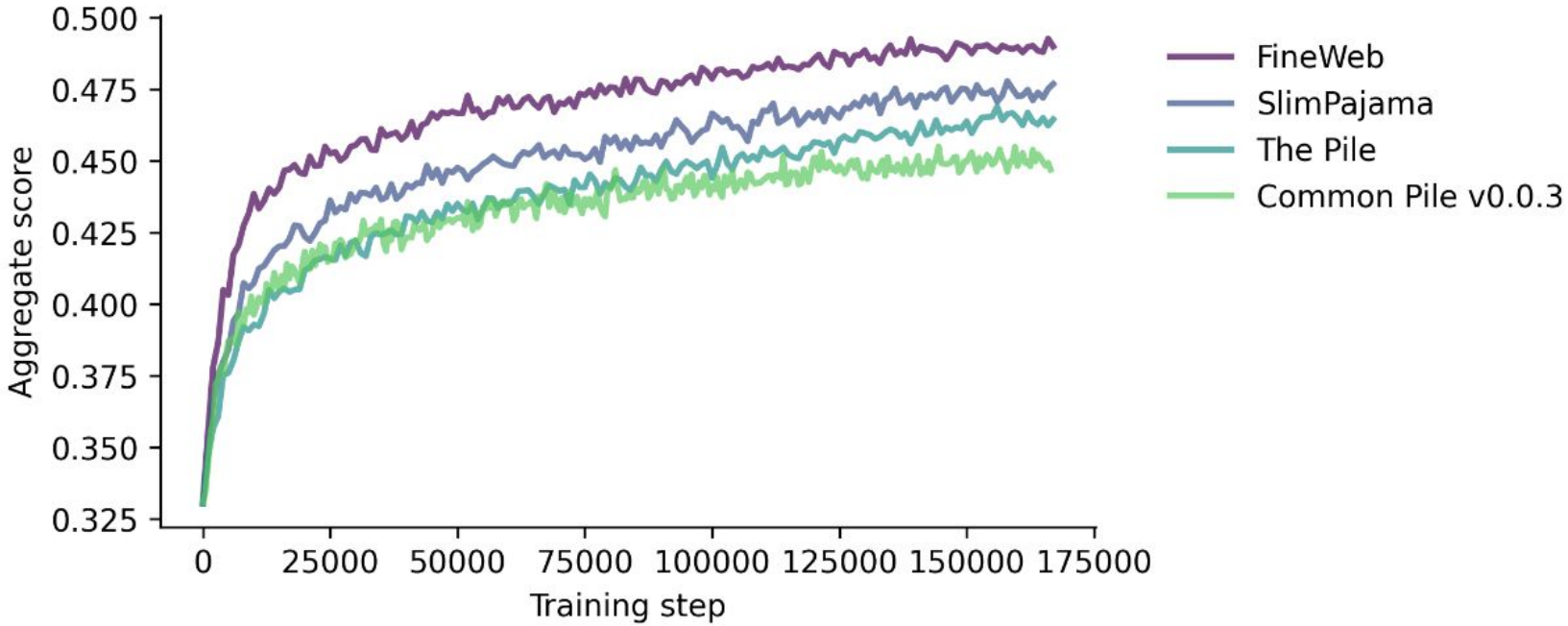
Proportion of licence types



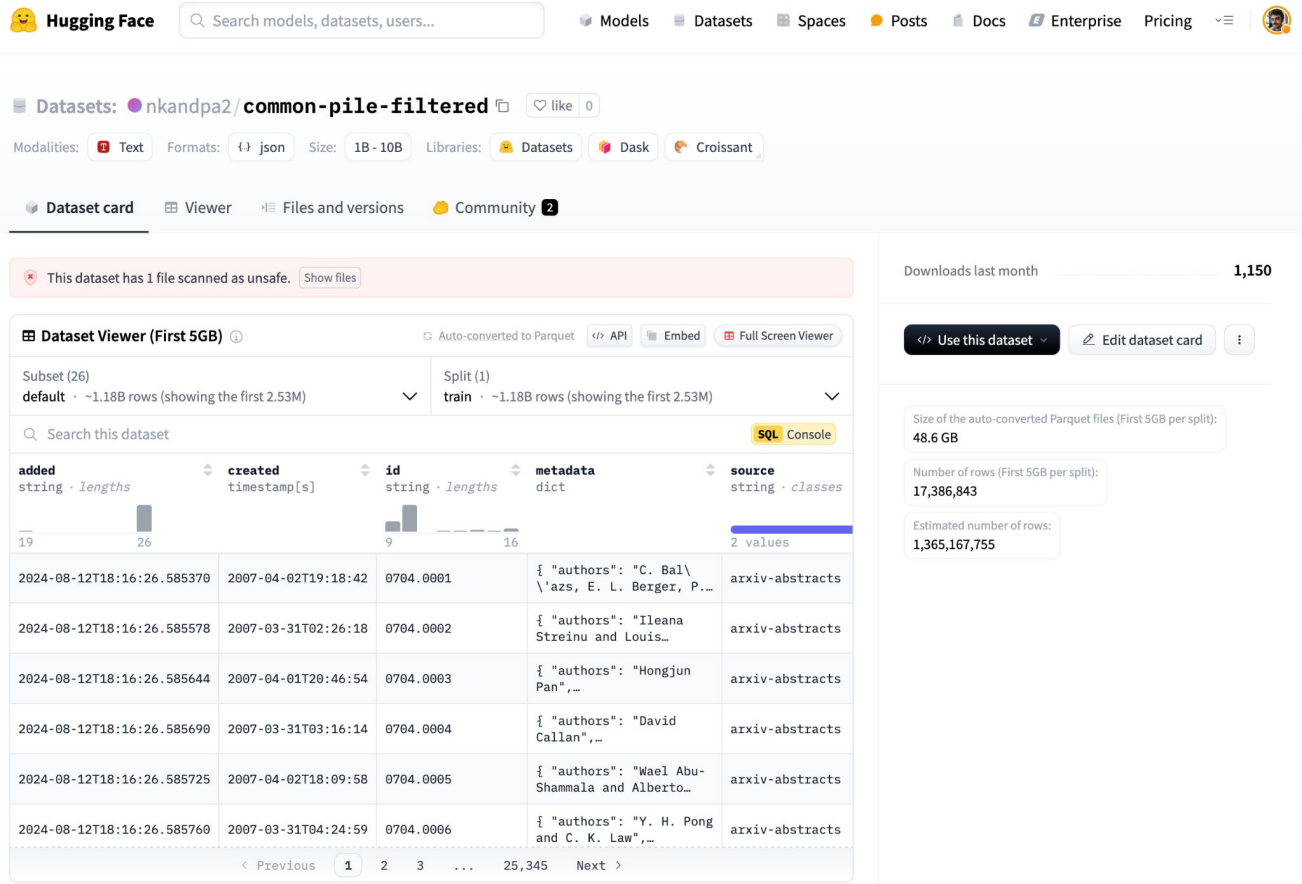
Proportion of source types



How it's going



Try it out!

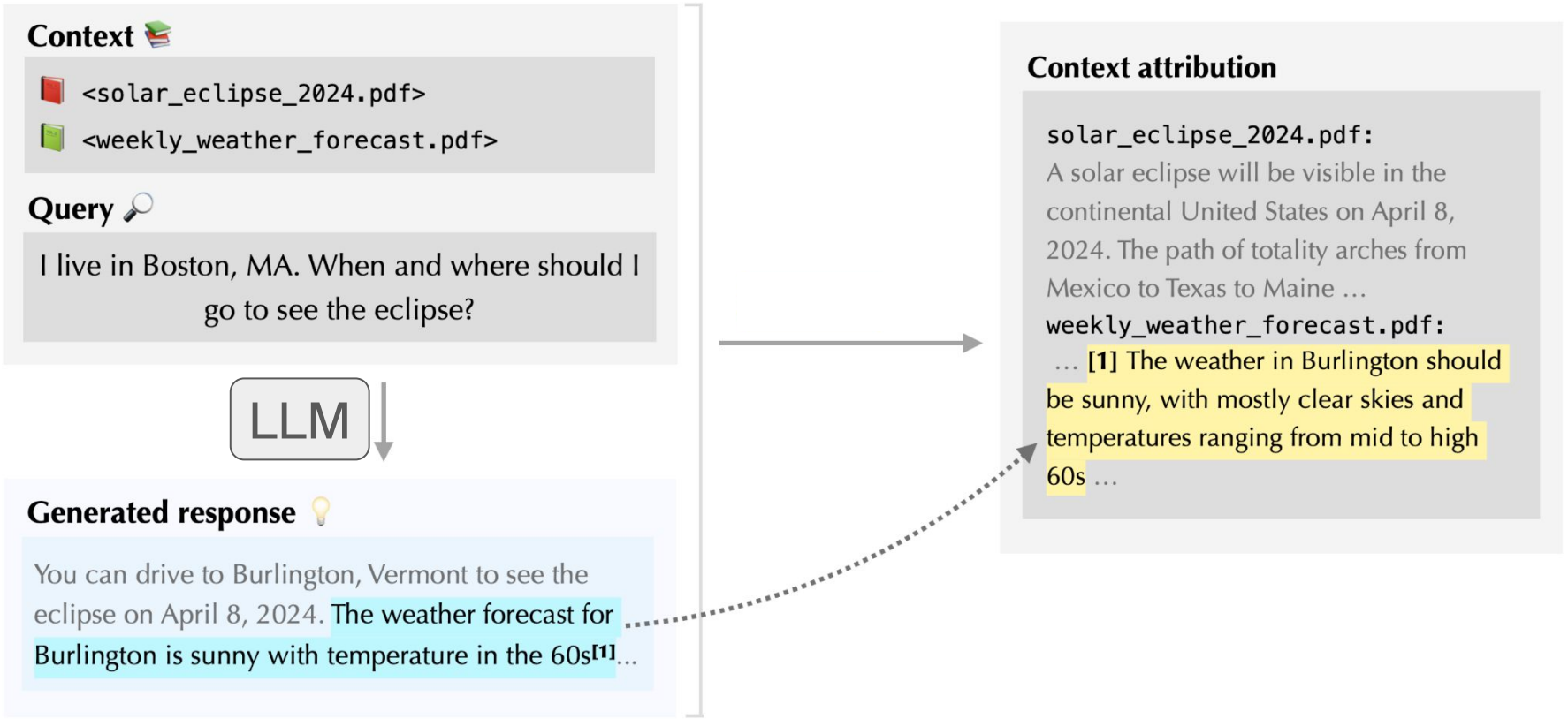


https://huggingface.co/datasets/nkandpa2/common-pile-filtered

Idea 2:

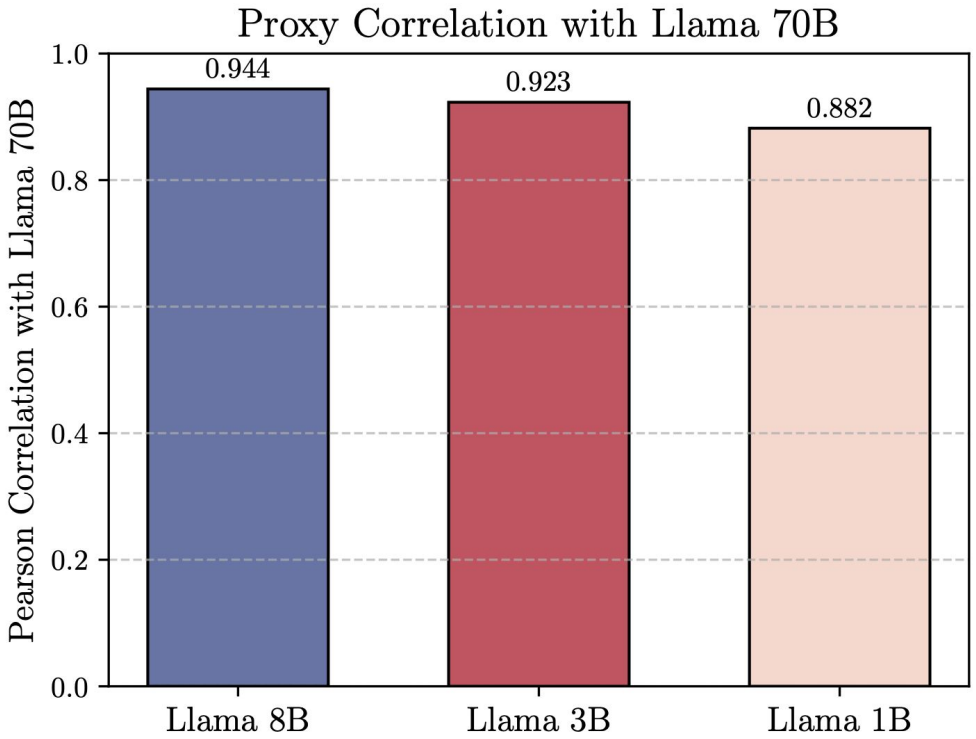
Attribute predictions back to training data

Context attribution finds influential spans in a model's input



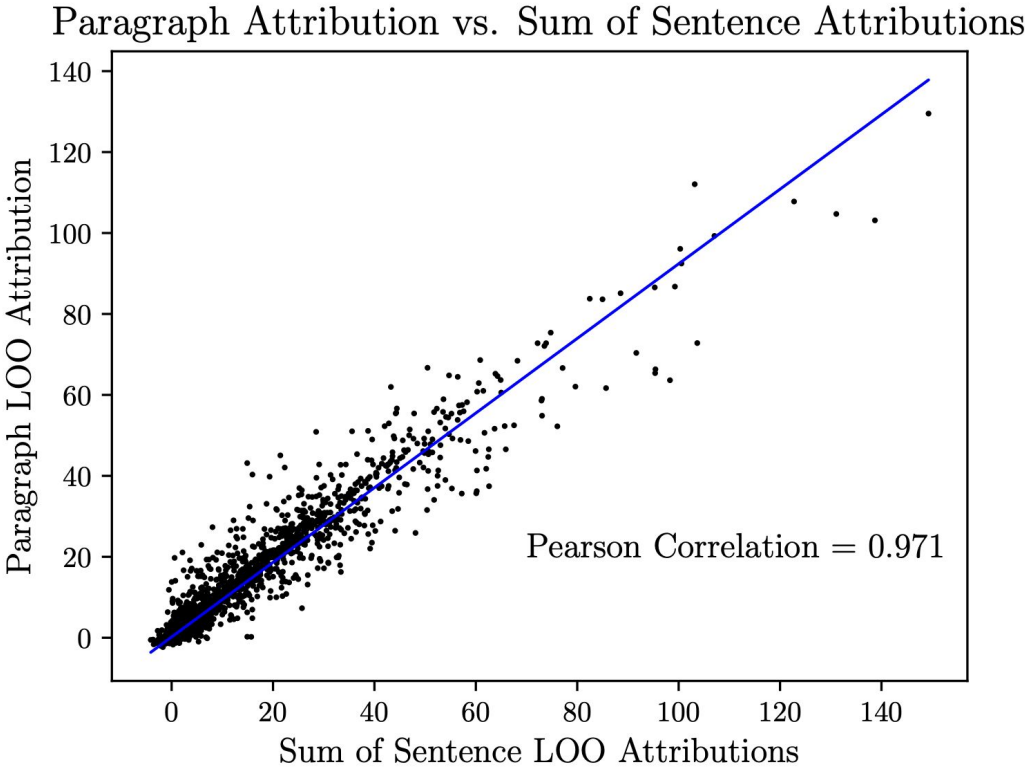
From "ContextCite: Attributing Model Generation to Context" by Cohen-Wang et al.

Observation 1: Smaller models provide similar attributions



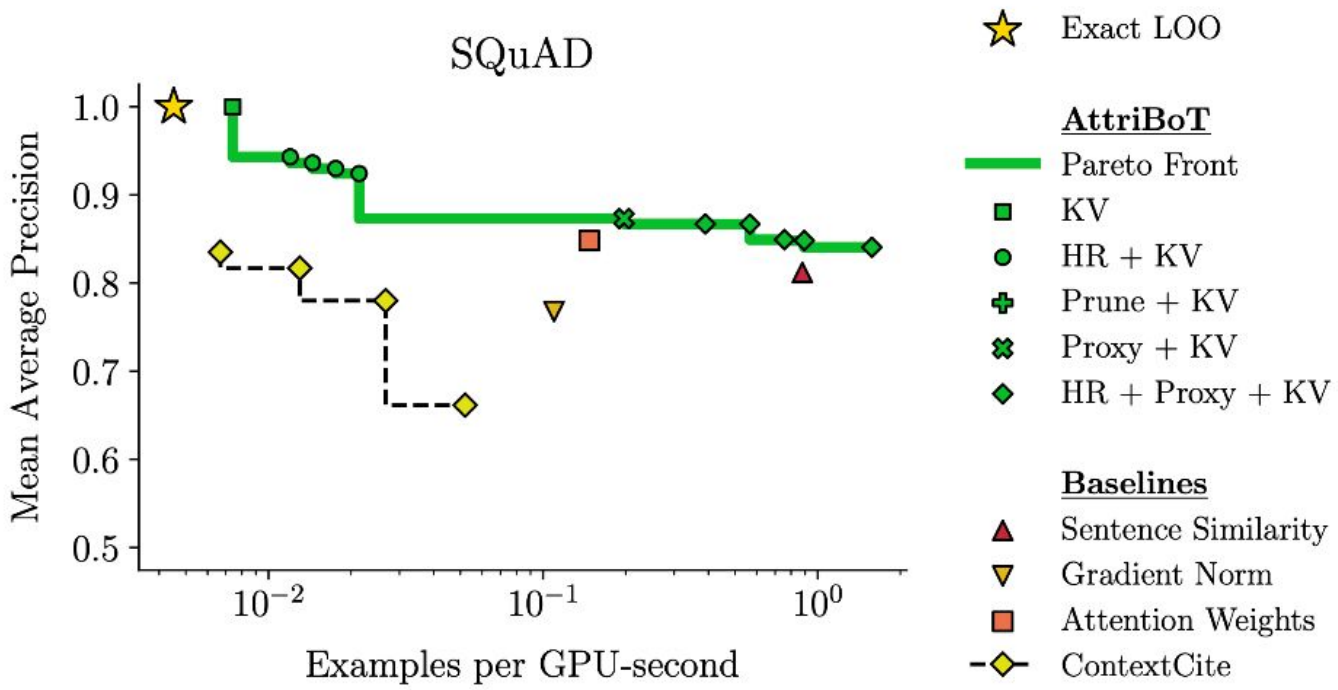
From "AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution" by Liu et al.

Observation 2: Attributions combine linearly



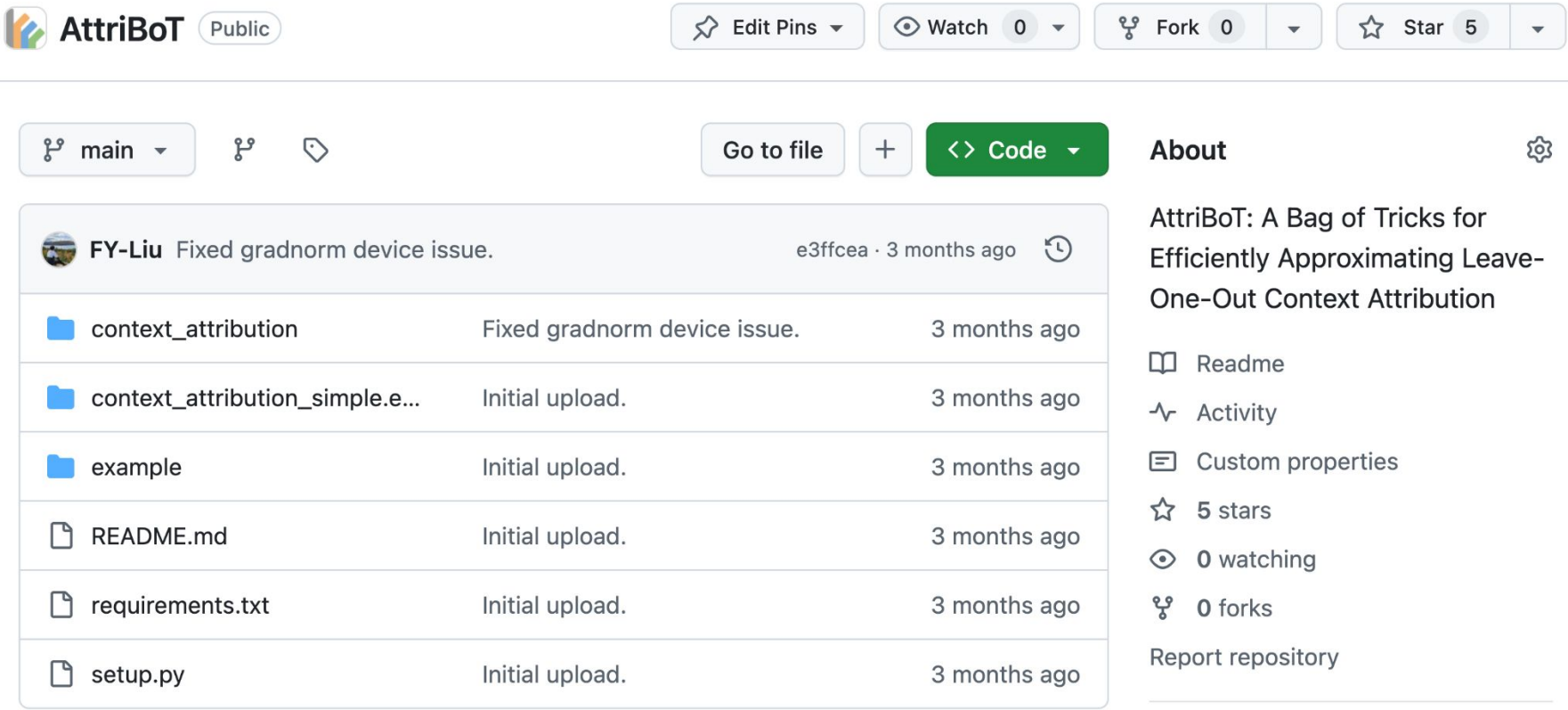
From "AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution" by Liu et al.

AttriBoT is pareto-optimal for context attribution



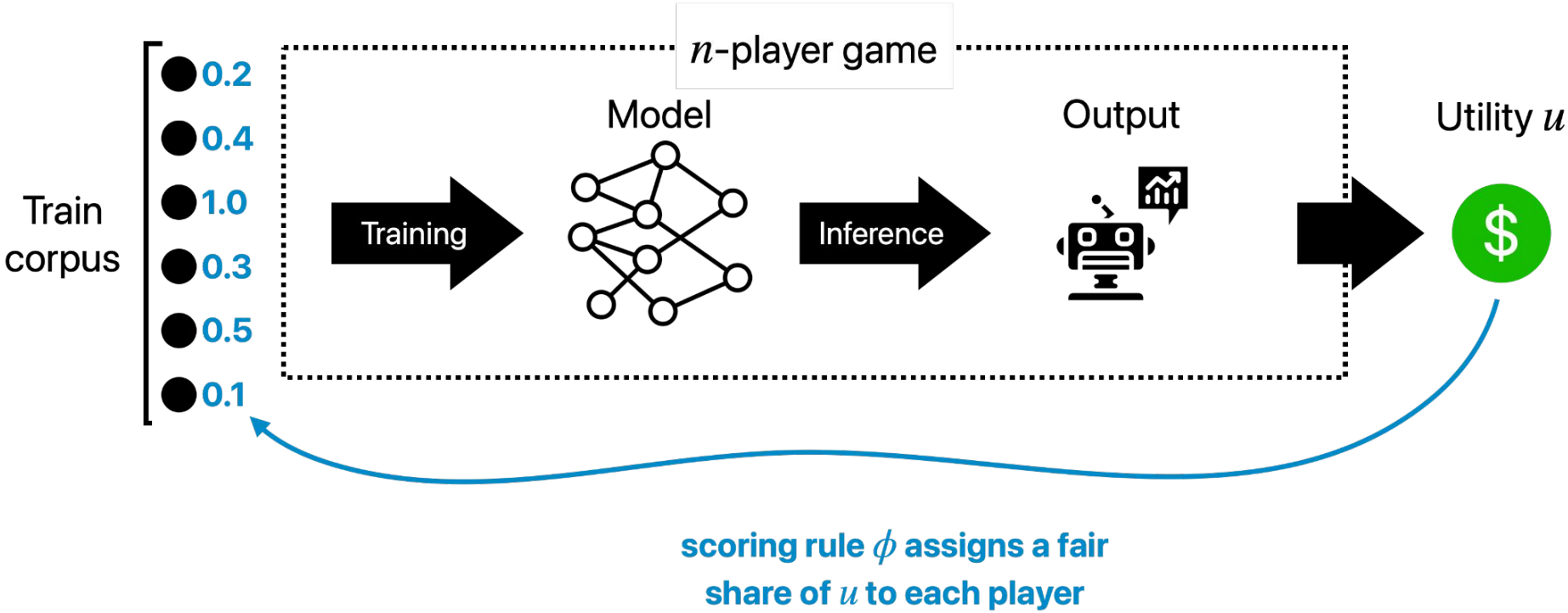
From "AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution" by Liu et al.

Try it out!



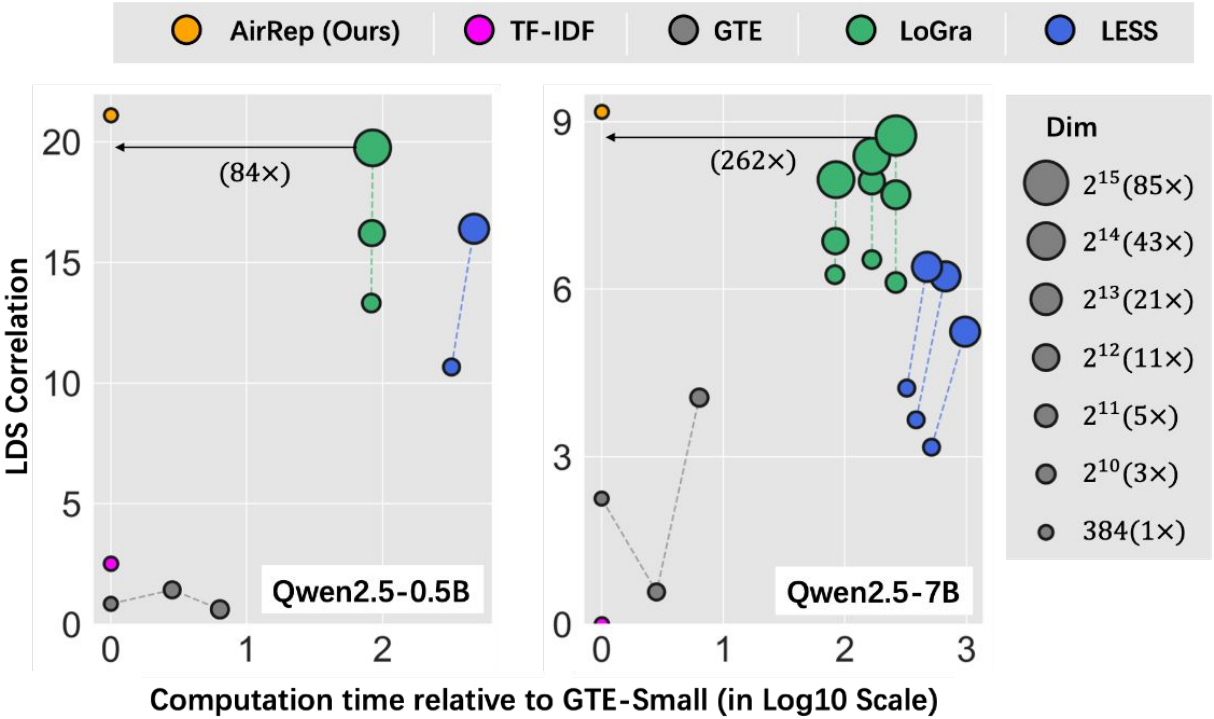
From https://github.com/r-three/AttriBoT

Measuring training data influence



From https://ml-data-tutorial.org/

AirRep is pareto-optimal for group data influence



From "Enhancing Training Data Attribution with Representational Optimization" by Sun et al.

Thanks.

Please give me feedback:

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

craffel@gmail.com