

Progress on a permissively licensed text dataset

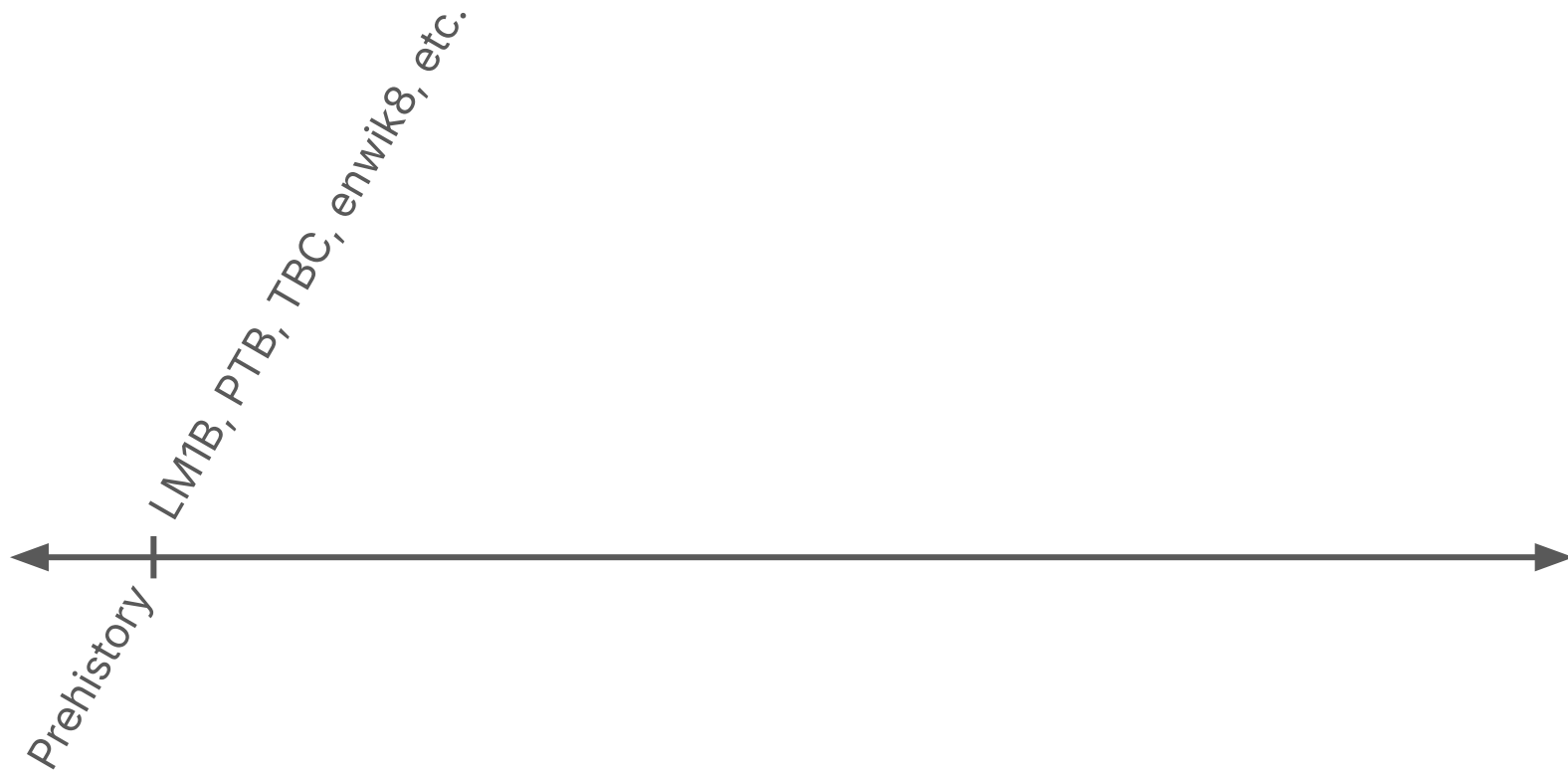
Colin Raffel



Language model training dataset timeline



Language model training dataset timeline



Early controversy with the Toronto Books Corpus

🕒 This article is more than **7 years old**

Google swallows 11,000 novels to improve AI's conversation

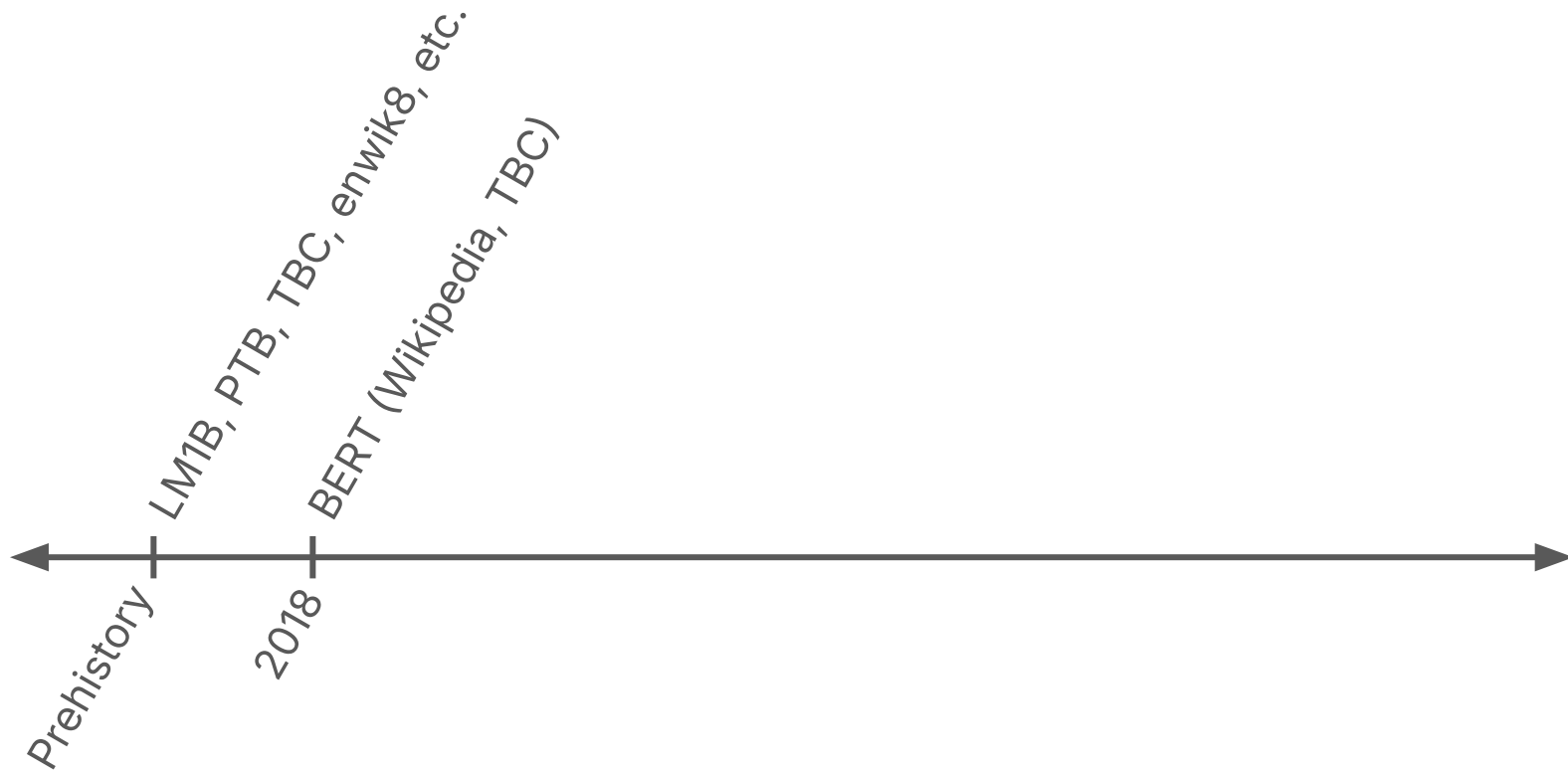
As writers learn that tech giant has processed their work without permission, the Authors Guild condemns 'blatantly commercial use of expressive authorship'

Richard Lea

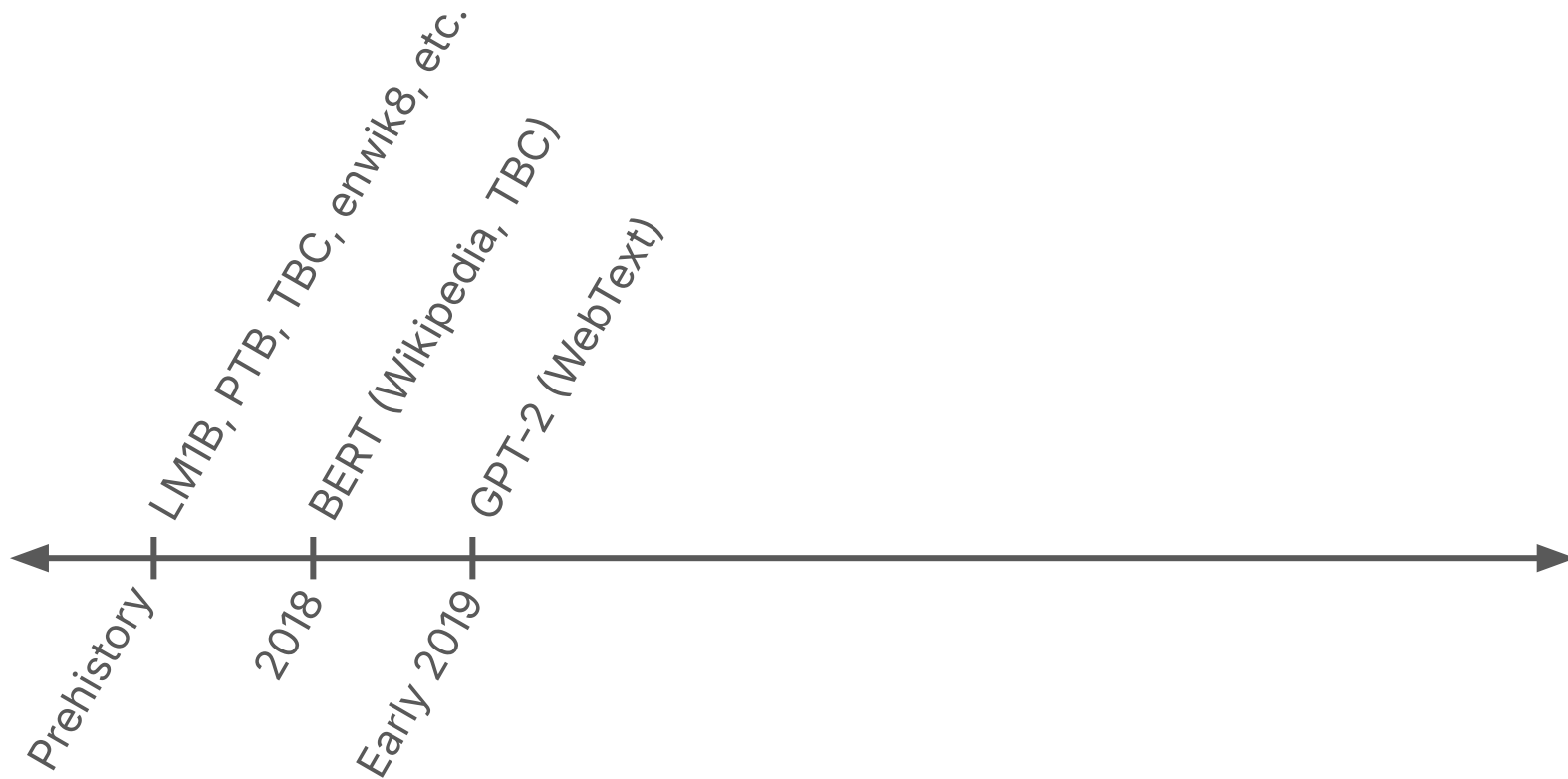
Wed 28 Sep 2016 10.00 BST

From <https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation>

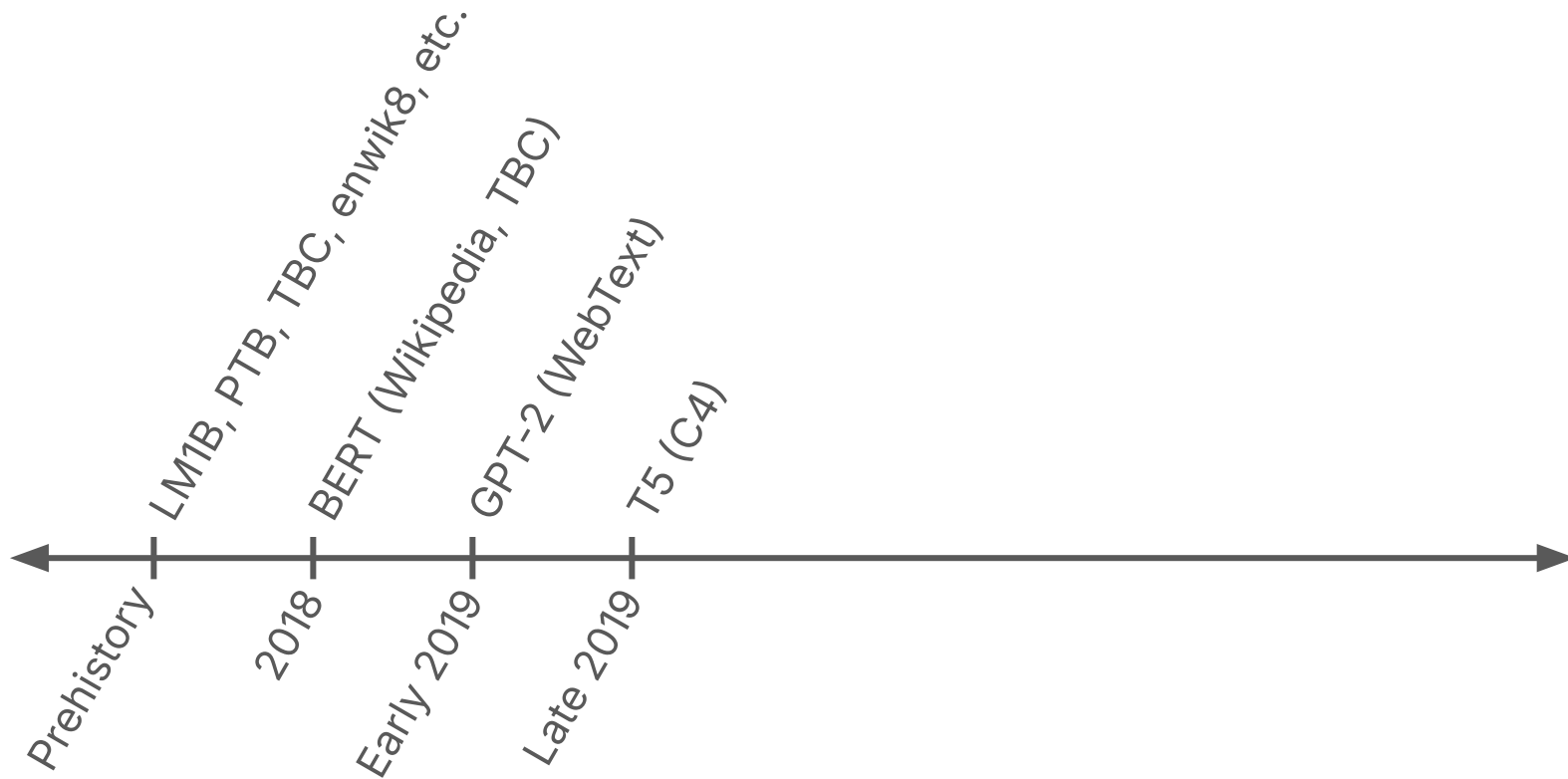
Language model training dataset timeline



Language model training dataset timeline



Language model training dataset timeline



Colossal Clean Crawled Corpus (C4)

Menu

Lemon

Introduction

The lemon, *Citrus Limon* (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a pH of around 2.2, giving it a sour taste.

Article

The origin of the lemon is unknown, though lemons are thought to have first grown in Assam (a region in northeast India), northern Burma or China. A genomic study of the lemon indicated it was a hybrid between bitter orange (sour orange) and citron.

Please enable JavaScript to use our site.

Home
Products
Shipping
Contact
FAQ

Dried Lemons, \$3.59/pound

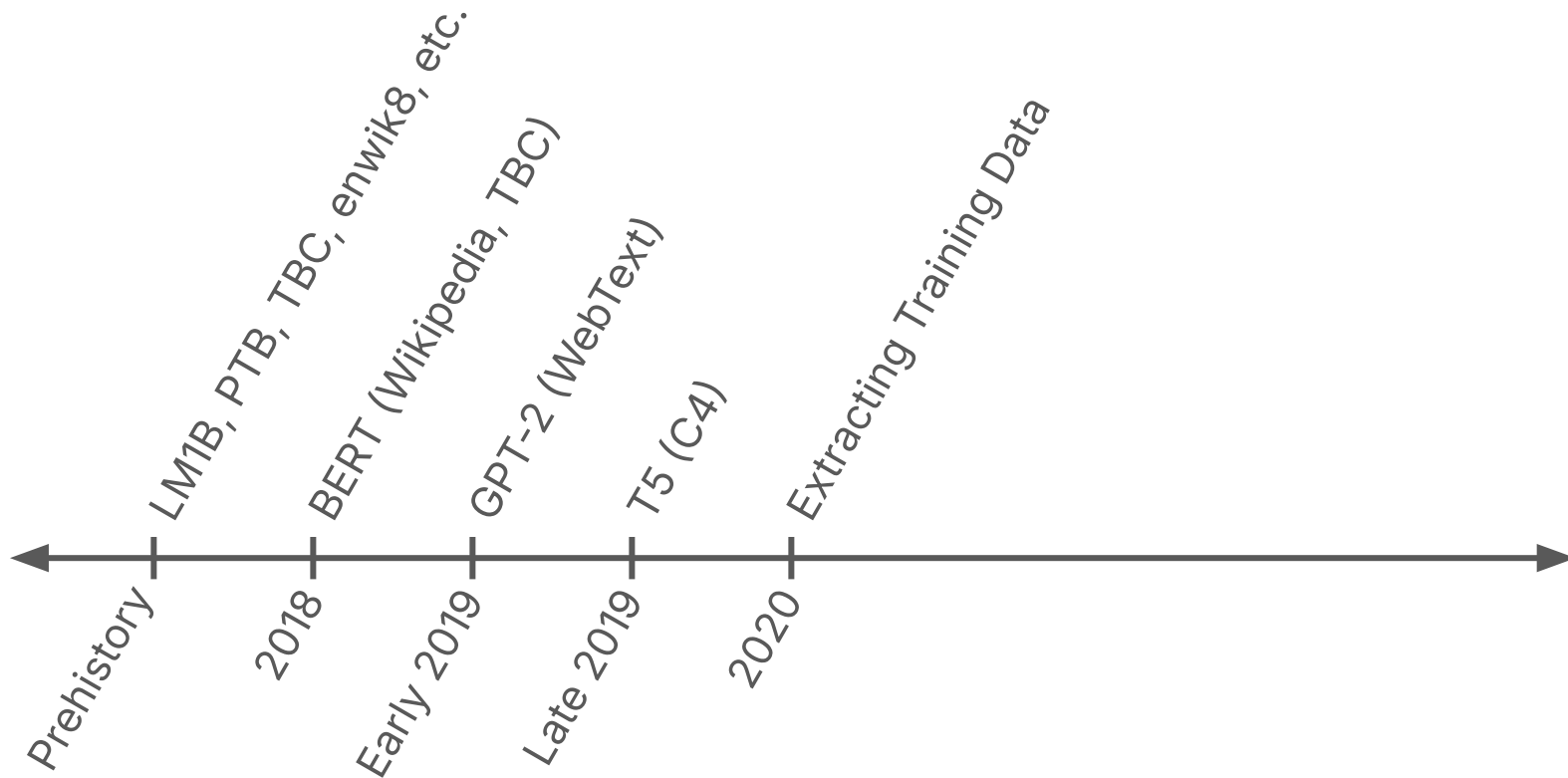
Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

The lemon, *Citrus Limon* (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a pH of around 2.2, giving it a sour taste.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur in tempus quam. In mollis et ante at consectetur.
Aliquam erat volutpat.
Donec at lacinia est.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.
Fusce quis blandit lectus.
Mauris at mauris a turpis tristique lacinia at nec ante.
Aenean in scelerisque tellus, a efficitur ipsum.
Integer justo enim, ornare vitae sem non, mollis fermentum lectus.
Mauris ultrices nisl at libero porta sodales in ac orci.

```
function Ball(r) {  
  this.radius = r;  
  this.area = pi * r ** 2;  
  this.show = function(){  
    drawCircle(r);  
  }  
}
```


Language model training dataset timeline

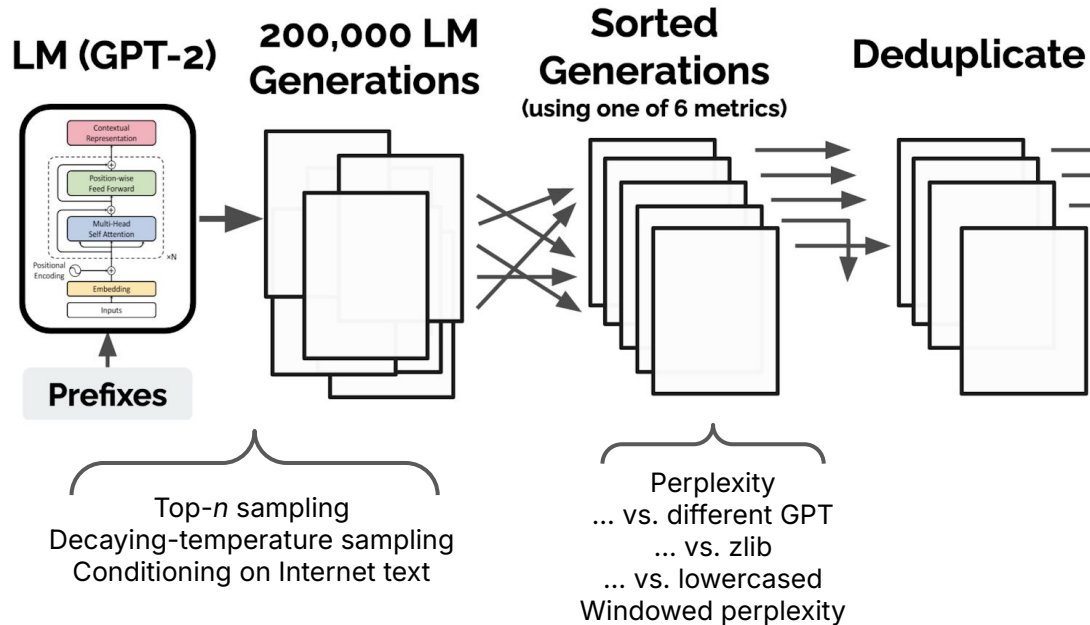


“Well-constructed AI systems generally do not regenerate, in any nontrivial portion, unaltered data from any particular work in their training corpus.”

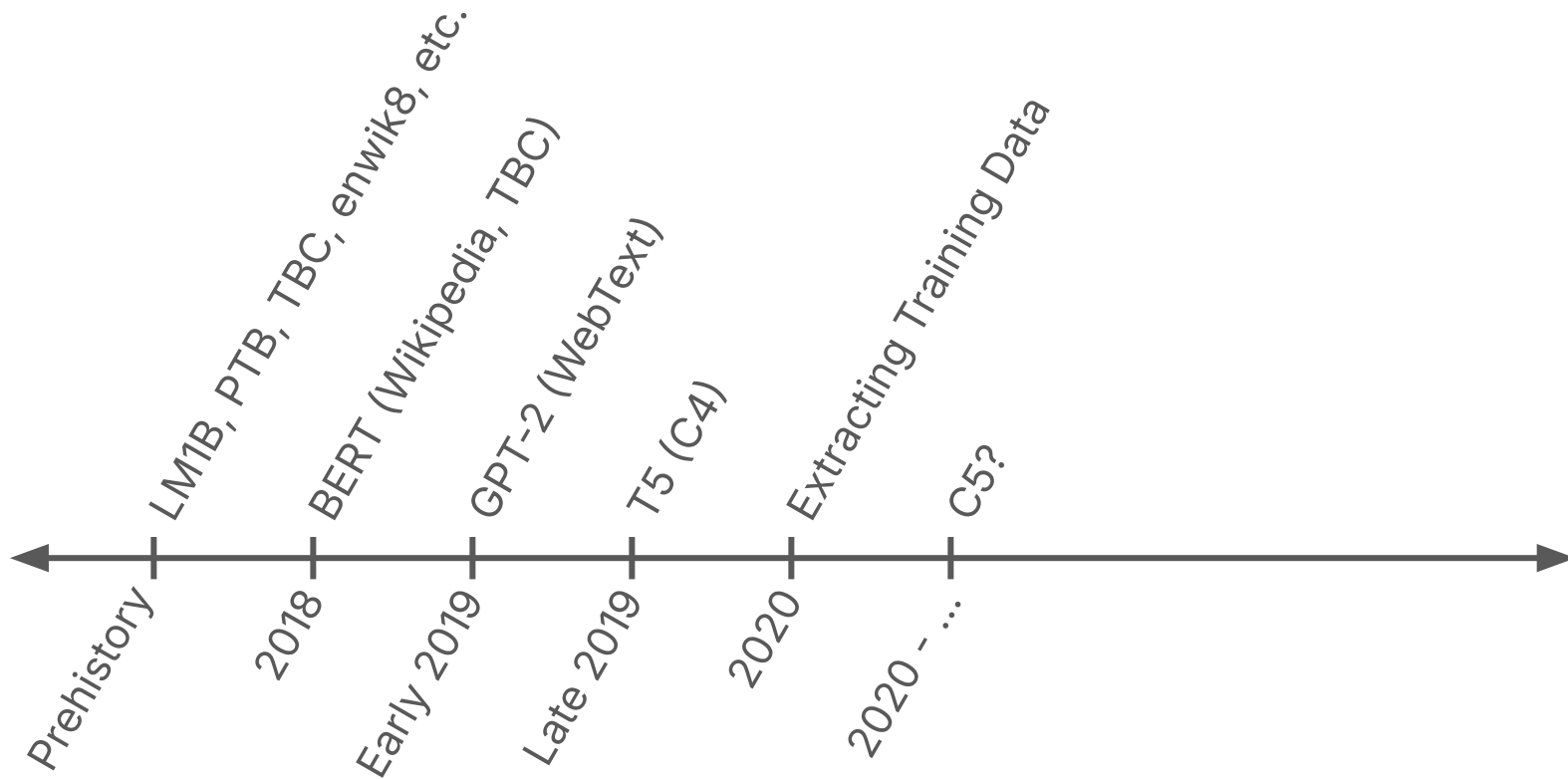
– [OpenAI](#)

Regenerating in a nontrivial portion unaltered data from the training set

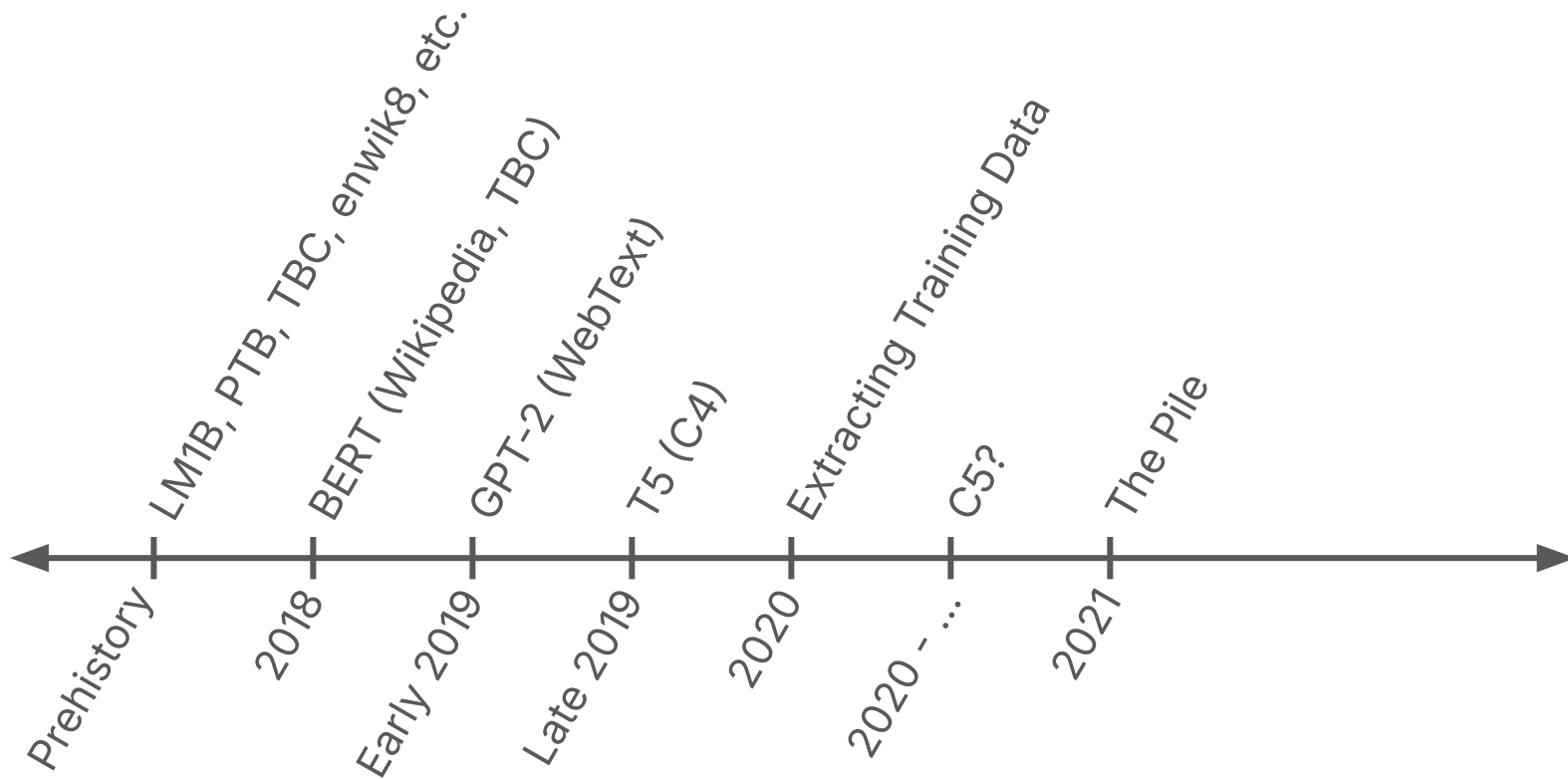
Training Data Extraction Attack



Language model training dataset timeline



Language model training dataset timeline

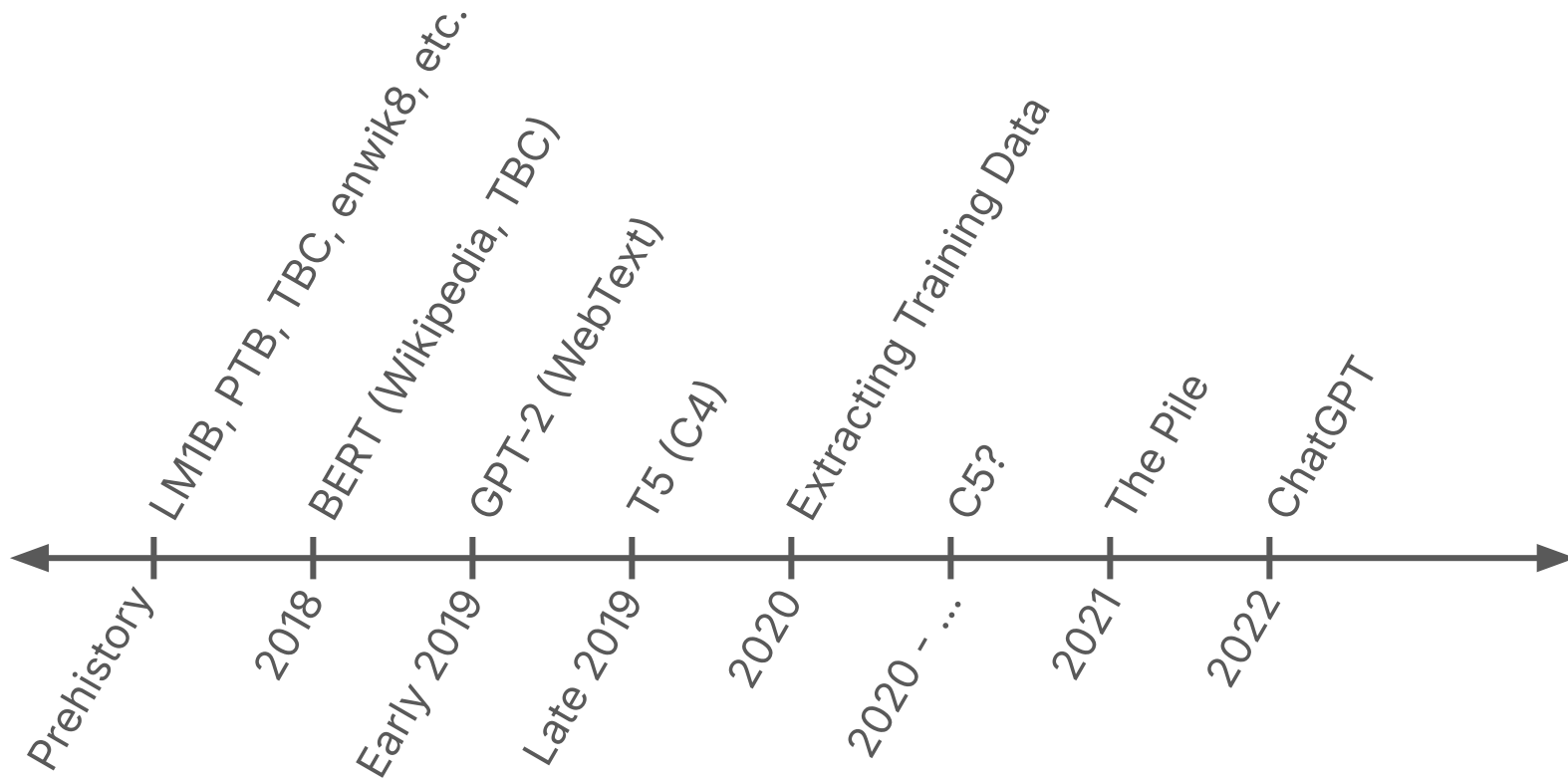


Sources in The Pile and their consent/licenses

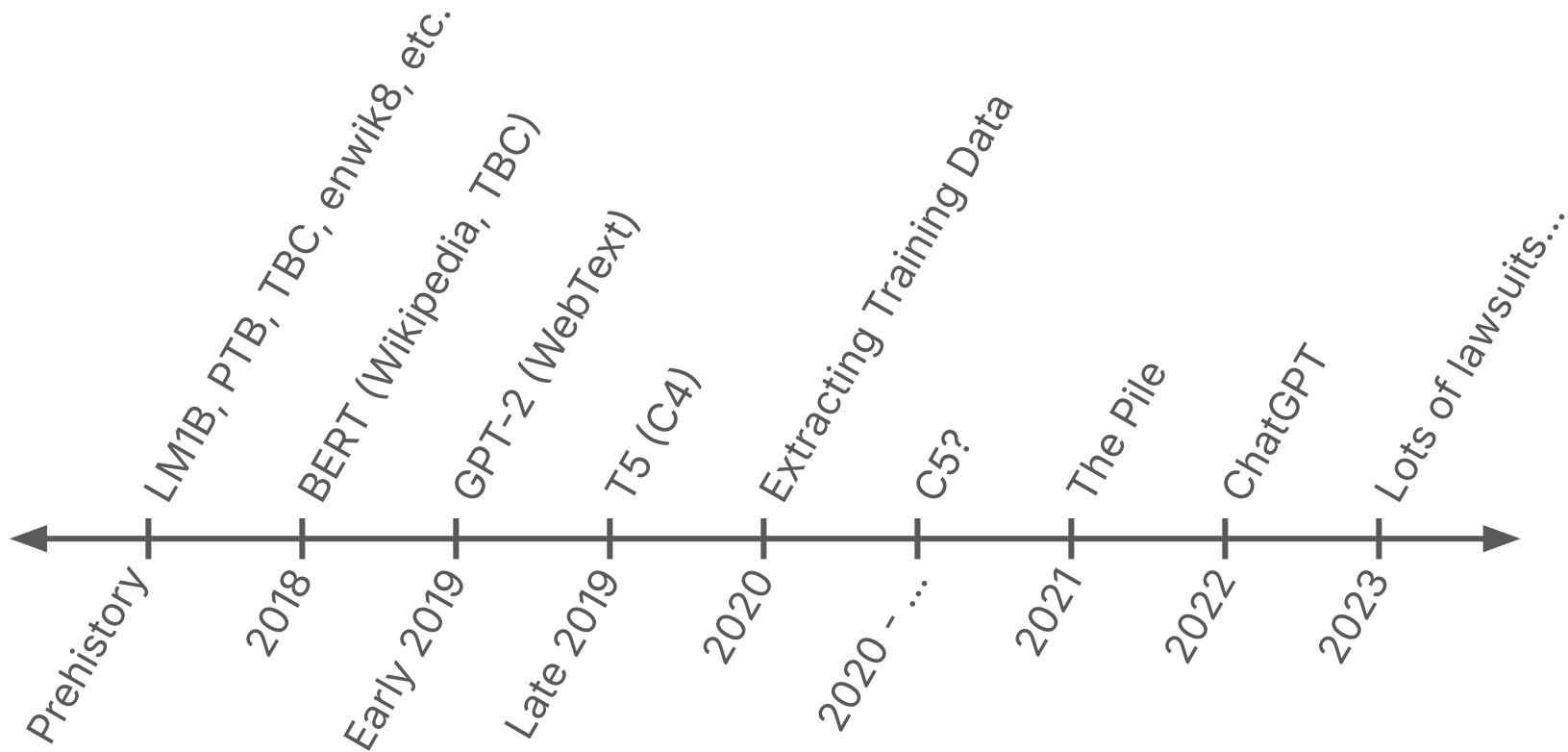
Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size	Public	ToS	Author
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB	✓	✓	
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB	✓	✓	✓
Books3 [†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB	✓		
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB	✓		
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB	✓	✓	✓
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB	✓	✓	
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB	✓	✓	✓
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB	✓	✓	✓
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB	✓	✓	✓
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB	✓	✓	✓
Gutenberg (PG-19) [†]	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB	✓	✓	
OpenSubtitles [†]	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB	✓		
Wikipedia (en) [†]	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB	✓	✓	✓
DM Mathematics [†]	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB	✓	✓	✓
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB	✓	✓	✓
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB	✓		
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB	✓	✓	✓
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB	✓	✓	
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB	✓		
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB	✓	✓	✓
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB	✓	✓	✓
Enron Emails [†]	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB	✓	✓	
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB			

From "The Pile: An 800GB Dataset of Diverse Text for Language Modeling" by Gao et al.

Language model training dataset timeline



Language model training dataset timeline



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](#)

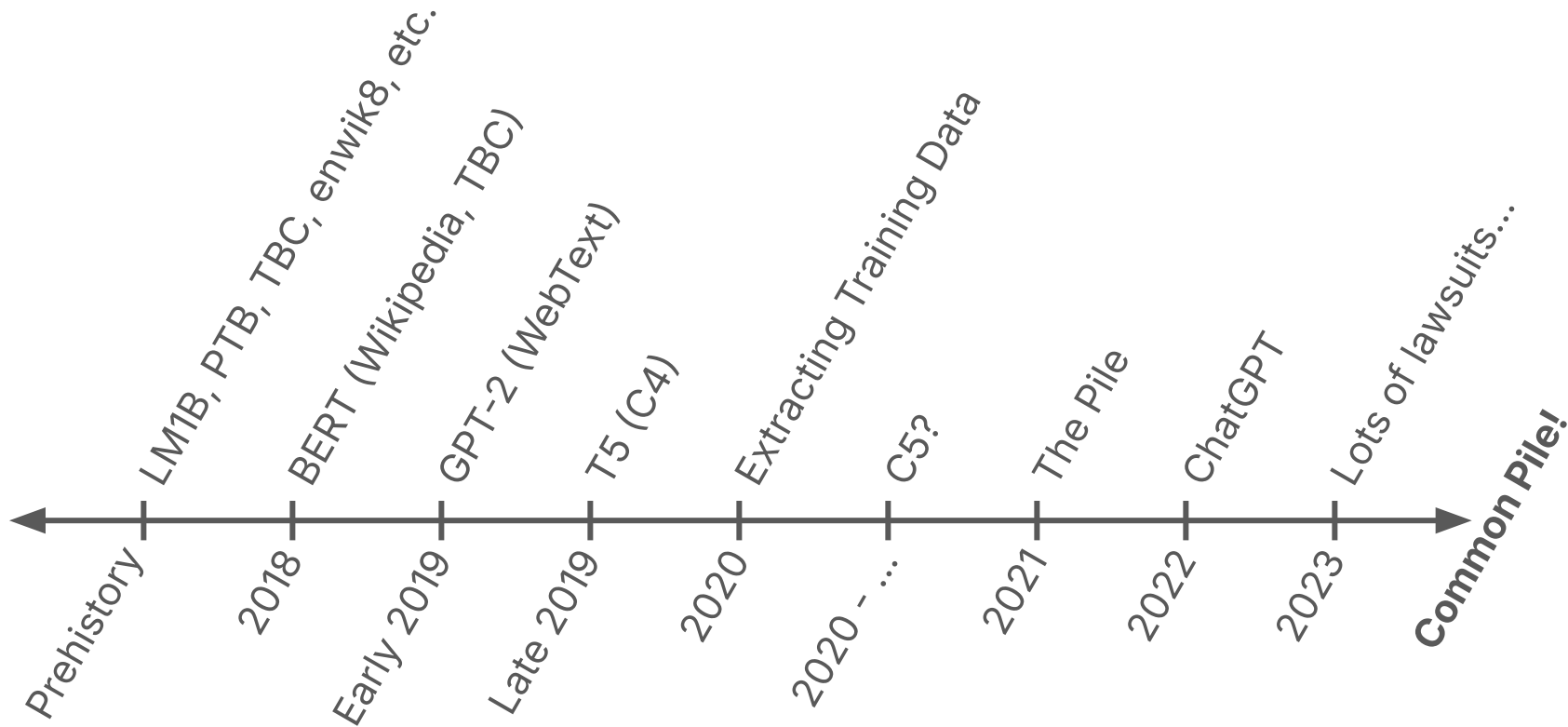
Media

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](#)

"... it would be impossible to train today's leading AI models without using copyrighted materials."

– [OpenAI](#)

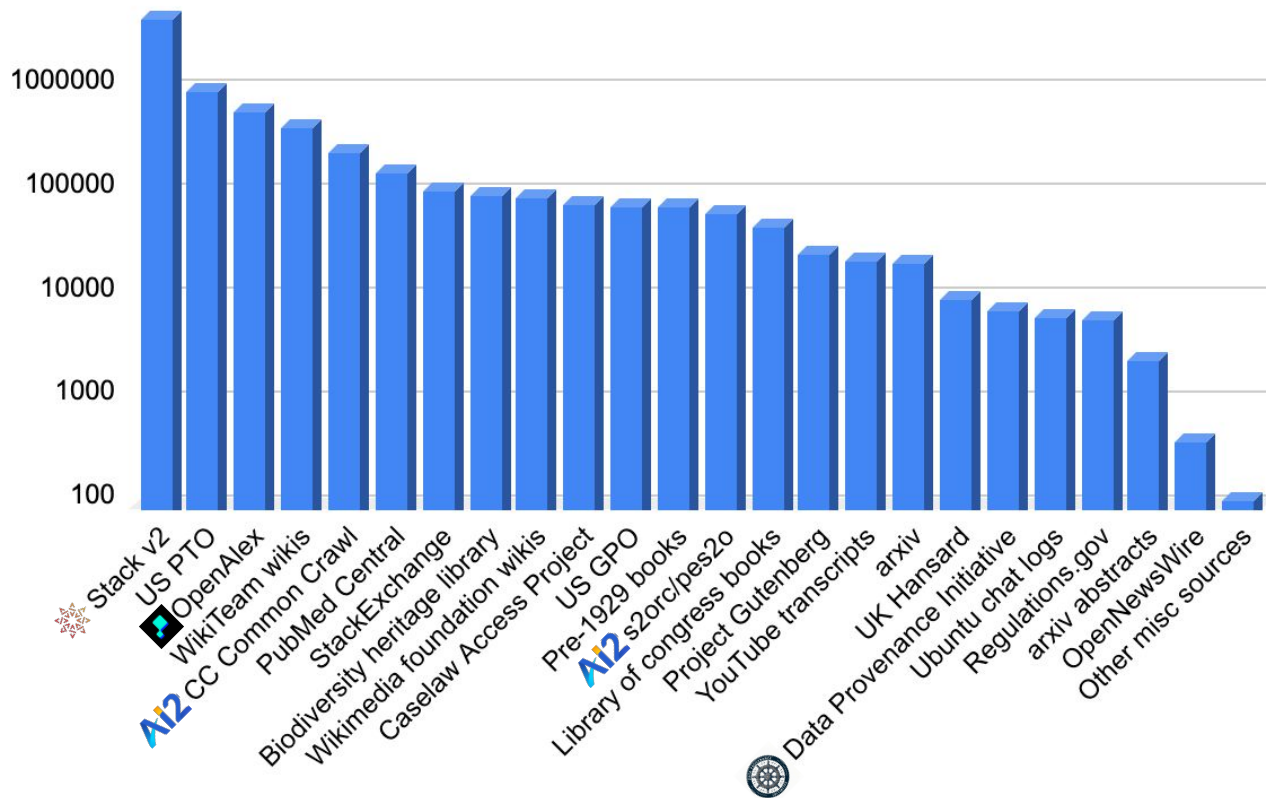
Language model training dataset timeline



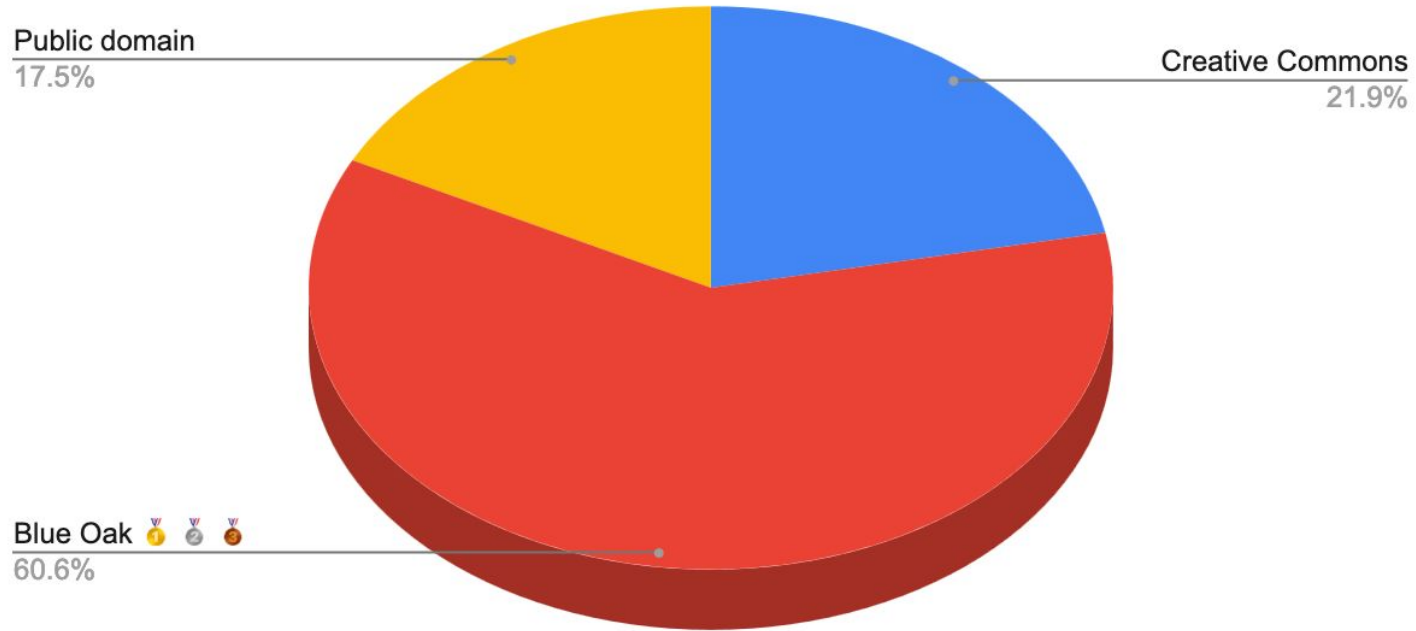
What is "permissively licensed" (to us)?

- 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

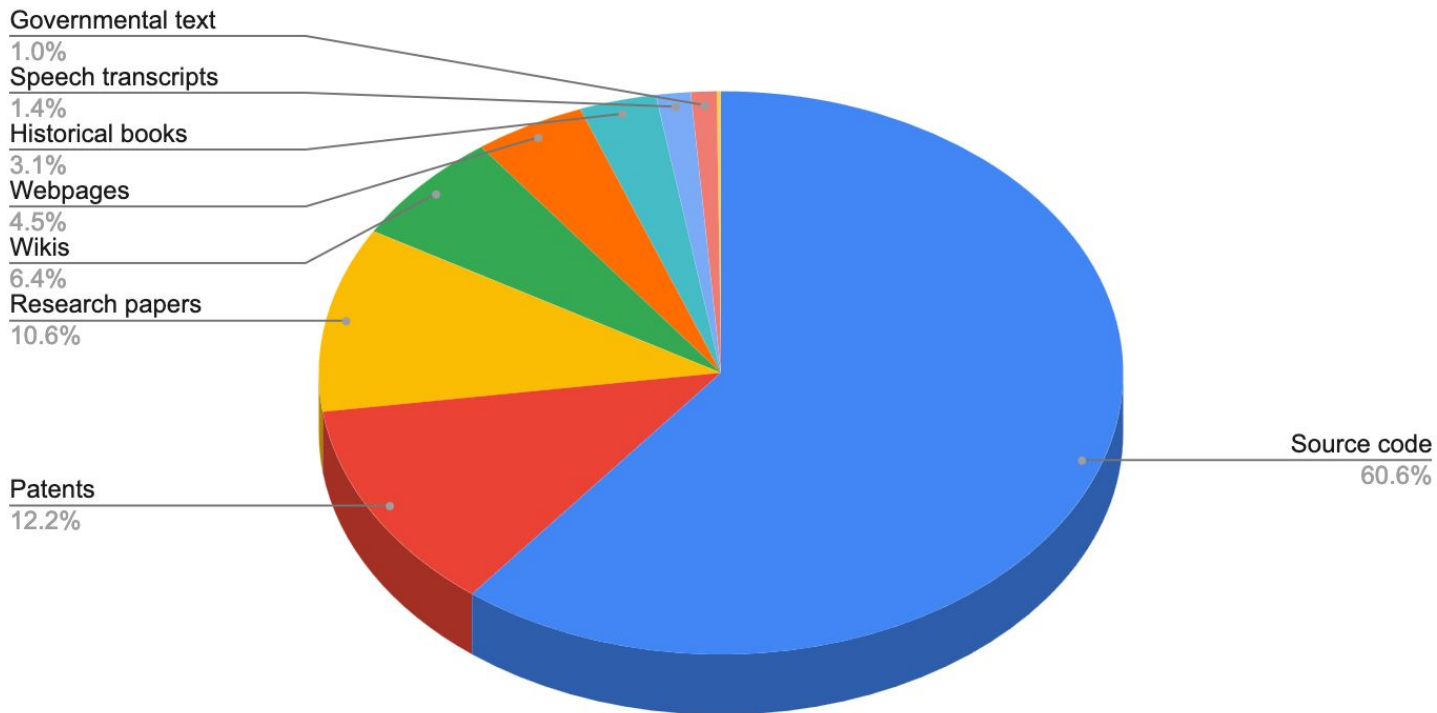
Raw source sizes (in UTF-8 bytes)



Proportion of licence types



Proportion of source types



Creative Commons Common Crawl

For the half-year to 30 June 2013, the IPKat's regular team is supplemented by contributions from guest bloggers Stefano Barazza, Matthias Lamping and Jeff John Roberts. Two of our regular Kats are currently on blogging sabbaticals. They are Birgit Clark and Catherine Lee. Friday, 17 August 2012

Taking it personally: patents, medicines and genetic markers

Purr-sonalised medicine

The IPKat is all in favour of medicine which, he thinks, can come in really handy -- even when you're unwell. He was therefore fascinated to learn of a recent discussion in Europe of a topic which has already exercised some of the finest minds of his American colleagues: the patenting of inventions relating to personalised medicine. Having heard about this from katfriend Suleman Ali (Holly IP), he is pleased to say that the latter was willingly persuaded to write a short note on the topic for the benefit of readers of this weblog. Here it is: "Is the EPO changing Its stance on personalised medicine inventions? Case law is an important means by which we know what is patentable at the European Patent Office (EPO). However, sometimes the EPO's view of what is patentable in an area changes before the case law does. This can sometimes be detected when Examiners start raising objections they would not have previously done. Clearly, applicants need to know about such changes as soon as possible so that they can revise their filing strategies and re-evaluate their expectations of the claims they are likely obtain. Meetings between the EPO and the epi (the professional institute for EPO attorneys) are very useful forums for obtaining 'inside information' about the EPO's thinking which is not yet apparent from the case law. The June 2012 issue of epi Information provides a report of such a meeting held on 10 November 2011 between the EPO and the biotech committee of the epi. Discussion item 8 is reported as follows: '8. Inventions in the area of pharmacogenomics This concerns cases which are based on a genetic marker to treat a disease, for example methylation profiles. It can involve a new patient group defined by an SNP. The EPO said that often the claims can lack novelty ...

US Patent and Trademark Office

A golf club carrier comprising a substantially rectangular frame member having a pair of wheel assemblies mounted thereon which extend toward the rear of the carrier. A club receiving container is permanently mounted on the frame and extends toward the rear thereof. The container is comprised of a, golf iron receiving compartment in the rear portion thereof and a storage compartment (for jackets, shoes, balls, etc.) in the front portion thereof. The golf iron compartment is adapted to receive golf irons therein with the heads thereof supported at the bottom of the compartment and with the shafts extending outwardly from the open top of the compartment to permit easy removal and insertion thereof. The frame member has an offset portion at the lower end thereof which provides a space for mounting a plurality of golf wood supporting pockets which are fastened to the bottom panel of the storage compartment. The golf wood pockets extend downwardly and rearwardly under the container and are adapted to receive the heads of golf Woods with the toes thereof pointed downwardly and rearwardly with respect to the container. With the wood heads tucked under the container, the adjacent frame portions provide a protective bumper to prevent denting and scratching of the wood club heads. The heads of the woods and irons are positioned below the level of the center points of the wheels to provide a low center of gravity for improved stability both when the carrier is at rest as well as when it is being pulled along the ground in use.

BACKGROUND OF THE INVENTION

Field of the invention
This invention relates to golf club carriers and more particularly to a golf club carrier wherein the containing and supporting means for the golf clubs is permanently mounted on the carrier.

Description of the prior art
The most pertinent prior art known to the applicant are US. Pat. 2,858,140 and 2,985,462. In the prior art patents referred to, the golf woods are mounted on the front portion of the carrier just as in the present application. However, in the prior art the wood clubs are supported in pockets fastened to the front portion of the club container which pockets extend forwardly from ...

Biodiversity Heritage Library

On the Withering of Plants

The Snowdrop is thus extinguished before the Crocus, and the Crocus before the after flowers. The scene must never be vacant, the old must remain with us till the new is well unfolded ; but we care little for the last lingering blossoms, and even if they were as lovely as ever, they would remain as a thing of a bygone day, in which our interest has ceased. Now if there were no withering, and the petals continued perfect till they fell from the stalk, a flower would contrast with its successors at a great disadvantage – we should feel that it was being outshone by them. But Nature will not permit her favourites to be dishonoured in this way, and she quietly withdraws them from the rivalry. When we have seen them as long as she thinks good to permit, she lays their beauty waste. But before this is done, a close observer will notice that the plant's most subtle and exquisite attractions have been stolen away imperceptibly, so that even whilst there is no sign of actual decay, the power of enchantment is lost, and that which finally palls upon our memory is not the flower, but the flower robbed of its soul, a mere copy of the great original masterpiece. And to carry out this

StackExchange

Why is Naruto hated in the beginning?

Why is Naruto hated so much when the village had so many Jinchurikis in the past? Didn't they know the risks of having a tailed beast so close by? My memory is vague but wasn't there an agreement that nobody mentions to Naruto that his body is being used as the fox's vessel? That might have been a factor in people not talking to him completely.

Hey welcome to A&M. How far are you already in the series? This will be explained fairly well throughout the series, and explanations might very well spoil you. Actually, he was mainly hated because Kyubi attacked the village and caused death of the 4th Hokage. Because he was its vessel, the hate of the villagers for Kyubi has been transferred to Naruto himself. First off all, they didn't hate him, but rather, they were afraid of him. This is because he had half of Kyubi seal into him. Don't forget one fact, that the Kyubi was used in two battles which much influence future of the village. He was summoned by Madara during his fight with Hashirama, and was released during the birth of Naruto by Obito. This was actually one of the most devastating attacks on Konoha so far. Many people died in this fight, for example, Iruka's parents, and The Hokage Minato. This event makes people of Konoha only see a symbol of destruction, death. I don't think he was hated per se, but I believe that they were afraid of him because he was being used as a vessel and they knew that. Plus, it could have been because he was just a kid without parents and living alone.

CourtListener

Wachenfeld, J. (dissenting). In *In re Woodworth*, 15 Fed. Supp. 291; affirmed, 85 F. 2d 50 (C. C. A. 2, 1936), the court held: "On principle, it cannot be doubted that when an attorney makes an agreement to prosecute a case for a fee contingent on success, and is disbarred before the fee is earned, he may not collect compensation from his client for the work done. The agreed fee he cannot have, because he has not performed his engagement and the contingency on which the compensation was to rest has not happened. Reasonable compensation in lieu of the fee he cannot have, because his inability to complete his contract has been brought about by his own wrongful ■conduct." I subscribe to this reasoning and conclusion and am therefore to affirm. Adopting this rule would not complicate or bring economic ■considerations into disciplinary proceedings nor would it defeat their purpose. It would, in my opinion, be an added incentive to professional conduct, which is foreign to disciplinary complaints. *530Admittedly, the plaintiff was disbarred because of his own wrongful act, and whether it was with reference to this particular case or not, the result, in my opinion, is the same. The penalty falls and he can no longer represent his client because of his wrongful conduct. The result of that misconduct should be uniform, not varying with the degree of culpability or its relationship to any particular case. "I-Iis inability to complete his contract has been brought about by his own wrongful conduct." I would affirm the judgment. For reversal – Chief Justice Vandebilt, and Justices Case, Heiiek, Olephant, Bueling and Ackekson – 6. For affirmance – Justice Wachenfeld – 1.

§.§ Training Protocols

In this section, we briefly review the training protocols that will be considered in this work (see [13] for detailed algorithmic tables). Throughout, we define the cross entropy between probability vectors \mathbf{a} and \mathbf{b} as $\phi(\mathbf{a}, \mathbf{b}) = -\sum_{l=1}^L a_l \log b_l$. As a benchmark, with Independent Learning (IL), each learning model at device k is trained on the local training set \mathbb{D}_k by using Stochastic Gradient Descent (SGD) with step size $\alpha > 0$ on the cross-entropy loss (see, e.g., [16]). With Federated Learning (FL) [1], at each global iteration i , each device k follows IL within the local training phase, and then it transmits the update $\Delta \mathbf{w}_{i,k}$ of the local weight vector $\mathbf{w}_{i,k}$ to the PS during the information exchange phase. The PS computes the average update $\Delta \mathbf{w}_i = 1/K \sum_{k=1}^K \Delta \mathbf{w}_{i,k}$ with respect to the previous iteration. This is broadcast to all devices and used to update the initial weight vector for the local training phase in the next iteration.

With Federated Distillation (FD) [3], each device k , during the information exchange phase of any iteration i , transmits the average logit vector

$$\begin{aligned} \mathbf{s}_{i,t,k} &= \mathbb{E}_{\mathbf{c}, \mathbf{t} \sim \mathbb{D}_k} [\mathbf{c} | \mathbf{t} = \mathbf{t}] \\ \mathbf{s}_{i,t,k} &= \mathbb{E}_{\mathbf{c}, \mathbf{t} \sim \mathbb{D}_k} [\mathbf{c} | \mathbf{t} = \mathbf{t}] \end{aligned}$$

Pre-1929 public domain books

MIST

THERE is a mist over this lake.

It shrouds the colors and the sounds as well ;

It is wrapped over the hills like a strong veil

It blurs the patterns that the pine-trees make, lace-
woven over the sky.

Old Sun, you can not pierce it;

As I look at you, you seem no more than a brightly-
cloudy glass sphere.

Little birds, your chirring is dull . . .

A cow-bell, clanking in the woods,
Has the muffled music of minor thirds.

Oh mist, you have lessened everything.

Even my longing is choked within my breast;

I can find no song for it.

Mediawiki wikis

Information theory

Information theory is the mathematical study of the quantification, storage, and communication of information. The field was originally established by the works of Harry Nyquist and Ralph Hartley, in the 1920s, and Claude Shannon in the 1940s. The field, in applied mathematics, is at the intersection of probability theory, statistics, computer science, statistical mechanics, information engineering, and electrical engineering.

A key measure in information theory is entropy. Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process. For example, identifying the outcome of a fair coin flip (with two equally likely outcomes) provides less information (lower entropy, less uncertainty) than specifying the outcome from a roll of a die (with six equally likely outcomes). Some other important measures in information theory are mutual information, channel capacity, error exponents, and relative entropy. Important sub-fields of information theory include source coding, algorithmic complexity theory, algorithmic information theory and information-theoretic security.

Applications of fundamental topics of information theory include source coding/data compression (e.g. for ZIP files), and channel coding/error detection and correction (e.g. for DSL). Its impact has been crucial to the success of the Voyager missions to deep space, the invention of the compact disc, the feasibility of mobile phones and the development of the Internet. The theory has also found applications in other areas, including statistical inference, cryptography, neurobiology, perception, linguistics, the evolution and function of molecular codes (bioinformatics), thermal physics, ...

Non-Mediawiki wikis

Man Vs. Society

Katniss's Society is riddled with problems. This is the exact reason why she is forced to enter the hunger games. The hunger Games serve as a reminder that the Capitol is in control. This is the most important conflict in the story.

Man Vs Man

Katniss is put in a situation where she is forced to fight and kill for her life against 23 other tributes until she is the only survivor left. She is not able to trust or count on any of the other tributes, for they might just be planning to kill her. The Hunger Games is all about the fights, the violence, and the deaths of these tributes as they struggle to return home alive.

Man Vs. Nature

Katniss is faced with many tough situations where she has to fend for herself in nature. When she is forced into the hunger games she must find water on her own, along with hunting food and nutrients in order to survive. It is clear that nature is not working with Katniss, it is working against her.

Man Vs. Self

She also fights to find hope in a hopeless world where everyone is out to get her. Another internal battle she faces is to keep her humanity despite the atrocities she's seen and even committed.

YouTube transcripts

This video will be about some rather strange integrals known as Borwine's integrals which have some rather odd properties. So the Borwine integrals are integrals of the sinc function so we recall that sinc is just \sin over x and sinc comes from the signals processing community. And its graph sort of looks like just decreasing oscillations. So its graph looks something like this. And its value here is 1. And it has the basic property that the integral from minus infinity to infinity of sinc of x dx is just π . This is a common exercise in either complex analysis or Fourier transform theory. And Borwein's integrals are a sort of generalization of this so this is the first one the next one you look at the integral from minus infinity to infinity of sinc of x times sinc of x over 3 dx and this is still π and then you can do the integral of x times sinc of x over 3 times sinc of x over 5, and this is still π . And you can go on like this up to the integral from minus infinity of sinc of x times all the way up to sinc of x over 13 and dx . And this is still π . But when you get to 15, it suddenly changes. It turns into this rather impressive number here. So here you notice this is very nearly equal to one.

US Government Publishing Office

Administration of Joseph R. Biden, Jr., 2024 June 10, 2024

Violent crime is dropping at record levels in America. It's good news for our families and our communities. Today the FBI released preliminary data collected from over 11,000 law enforcement agencies around the country showing that, in the first quarter of this year, murders decreased by 26 percent, robberies by almost 18 percent, and violent crime overall by 15 percent. These large decreases follow major reductions in crime in nearly every category in 2023, including one of the lowest rates for all violent crime in 50 years and significant declines in murder. This progress we're seeing is no accident. My administration is putting more cops on the beat, holding violent criminals accountable, and getting illegal guns off the street, and we are doing it in partnership with communities. As a result, Americans are safer today than when I took office. After we saw the largest increase in murders ever recorded during the previous administration, my administration got to work protecting the American people. My America Rescue Plan—which every Republican voted against—delivered \$15 billion to cities to hire and retain more cops and keep communities safe. I took on the gun lobby and signed the Bipartisan Safer Communities Act into law, the most significant gun violence legislation in nearly 30 years. But there is more to do. I will continue fighting for funding for 100,000 additional police officers, and crime prevention and community violence intervention programs. Every American deserves to feel safe in their community, which is why I will continue to invest in public safety. Categories: Statements by the President : Federal Bureau of Investigation Quarterly Uniform Crime Report for first quarter of 2024. Subjects: Community violence-interruption programs; Crime rates; Federal Bureau of Investigation; Gun violence, prevention efforts; Illegal firearms, antitrafficking efforts; State and local law enforcement, Federal support. DCPD Number: DCPD202400503. 1

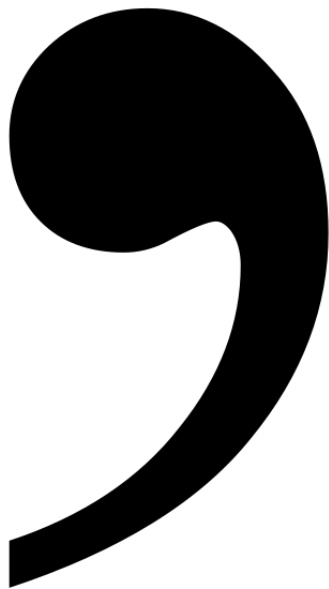
Hansard

[Stage Direction] Bill (No. G3) to amend the Franchise Act, 1898.—Hon. Mr. Fitzpatrick. [Stage Direction] The SOLICITOR GENERAL (Hon. Charles Fitzpatrick) asked for leave to introduce Bill (No. 64) to amend the Dominion Elections Act, 1900. He said : This Bill provides for amendments to the section respecting the posting of proclamations in the Territories. It also provides that in the case of the deposit required with the returning officer, a certified cheque on a bank will be available as deposit. Then it provides for a change in the form of the ballot to obviate difficulties arising in certain sections of Ontario where some of the voters marked their ballots outside the space opposite the name. It also provides that in the case of manhood suffrage voters, those who move from one district to another will not lose their votes, and makes provision that in the event of manhood suffrage registration lists having been prepared under the Dominion Act within the year previous to a by-election, they may be available for that new election, and do away with the necessity of making a new list, as is the case now with the lists in force in the provinces. Mr. MACLEAN.: Does the hon. gentleman intend to make any provision for Mr. Thornton getting the seat to which he has been elected, and for the reference of cases arising out of little imperfections in carrying out the law, to a county judge, where speedy justice can be obtained, as in the case of a recount and in municipal election cases ? The SOLICITOR GENERAL.: The particular election of West Durham is a matter which does not come within the purview of my powers, but within the powers of the House generally. The hon. gentleman's second point deserves consideration, and when the Bill comes up I will try to provide some short method of disposing of preliminary objections. Mr. BORDEN (Halifax).: All my hon. friend meant was whether the provision of the Bill with respect to a certified cheque would be made retroactive. The SOLICITOR GENERAL.: I would have to ask my hon. friend's opinion as to the propriety of such retroactive legislation. [Stage Direction] Motion agreed to. and Bill read the first time. Mr. THOMAS G.@: RODDICK (Montreal, St. Antoine) moved for leave to introduce Bill (No. 65) to ...

regulations.gov

Applicant: AEROLINEAS EJECUTIVAS, S.A. de C.V. Date Filed: September 12, 2000 Relief requested: Exemption from 49 USC section 41301 to permit the applicant to continue to conduct passenger charter operations between Mexico and the United States, and other passenger charter operations in accordance with 14 CFR Part 212, using small equipment. If renewal, date and citation of last action: November 18, 1999; in this Docket. Applicant representative(s): Lee A. Bauer, 202-331-3300 Responsive pleadings: None. DISPOSITION Action: Approved. Action date: November 22, 2000 Effective dates of authority granted: November 22, 2000, through November 22, 2001. Basis for approval (bilateral agreement/reciprocity): United States-Mexico Air Transport Services Agreement of August 15, 1960, as amended and extended (Agreement). Except to the extent exempted/waived, this authority is subject to the terms, conditions, and limitations indicated: X Standard exemption conditions. Special conditions/Partial grant/Denial basis/Remarks: In the conduct of these operations, the carrier must adhere to all applicable provisions of the U.S.-Mexico Agreement. In the conduct of these operations, the carrier may only use aircraft capable of carrying no more than 60 passengers and having a maximum payload capacity of no more than 18,000 pounds (small equipment). The above grant includes authority to conduct Third and Fourth Freedom charter operations. While we have subjected, consistent with the provisions of the Agreement, Mexican carriers conducting charter operations with large aircraft to prior approval of their Third and Fourth Freedom charters (see Order 92-2-7 at 5), we determined that a Third/Fourth Freedom prior-approval requirement was not necessary on public interest grounds in the case of this carrier, since it will be conducting these operations solely with small aircraft. (Other charter operations to/from the United States under this authority, however, are subject to prior approval under 14 CFR Part 212.) Further, we are continuing to allow Mexican carriers conducting passenger charters using small equipment to make stopovers in the United States in the conduct of such operations. Action taken by: Paul L. Gretch, Director ...

From raw text to training dataset



Filtering pipeline and content removed by filtering/deduplication

Source	Filtering						
	LangID	Toxicity	PII	Doc len	Line len	LM/OCR	Dedupe
arxiv	Y	N	Y	N	N	N	N
arxiv abstracts	N	N	Y	N	N	N	N
Biodiversity heritage library	Y	N	N	Y	N	Y	N
CL / Caselaw Access Project	N	Y	Y	Y	N	N	N
CC Common Crawl	Y	Y	Y	Y	Y	N	Y
Data Provenance Initiative data	N	N	N	N	N	N	Y
Foodista	Y	N	Y	Y	N	N	Y
UK Hansard	Y	N	Y	N	Y	N	N
Library of congress books	N	Y	N	N	N	Y	Y
Project Gutenberg	Y	N	N	N	N	Y	Y
OpenAlex	N	N	Y	N	N	N	Y
OpenNewsWire	Y	N	Y	Y	N	N	Y
Public Domain Review	N	N	Y	Y	N	N	N
PubMed Central	N	N	Y	Y	Y	N	N
Pre-1929 books	N	Y	N	N	N	Y	Y
Python PEPs	N	N	Y	N	N	N	N
Regulations.gov	N	N	Y	Y	N	N	N
s2orc/pes2o	N	N	Y	N	N	N	N
StackExchange	Y	N	Y	N	N	N	N
Stack v2	N	N	N	N	N	N	N
Ubuntu chat logs	Y	Y	Y	Y	N	N	N
US Government Publishing Office	N	N	N	N	N	N	Y
US Patent and Trademark Office	N	N	Y	Y	N	Y	N
WikiTeam wikis	Y	Y	Y	Y	N	N	Y
Wikimedia foundation wikis	Y	N	Y	Y	N	N	N
YouTube transcripts	Y	Y	Y	Y	N	N	N

Source	Percentage removed from...	
	Filtering	Deduplication
Total	50%	9%
Stack v2	72%	0%
US PTO	5%	1%
OpenAlex	0%	40%
WikiTeam wikis	83%	76%
CC Common Crawl	22%	9%
s2orc/pes2o	0%	2%
PubMed Central	7%	0%
StackExchange	14%	0%
Biodiversity heritage library	55%	3%
Wikimedia foundation wikis	10%	4%
Caselaw Access Project	0%	1%
US GPO	0%	50%
Pre-1929 books	8%	3%
Library of congress books	7%	0%
Project Gutenberg	21%	2%
YouTube transcripts	13%	0%
arxiv	6%	0%
UK Hansard	3%	1%
Data Provenance Initiative	0%	18%
Ubuntu chat logs	16%	0%
Regulations.gov	0%	16%
arxiv abstracts	0%	2%
OpenNewsWire	15%	3%
Other misc sources	4%	1%

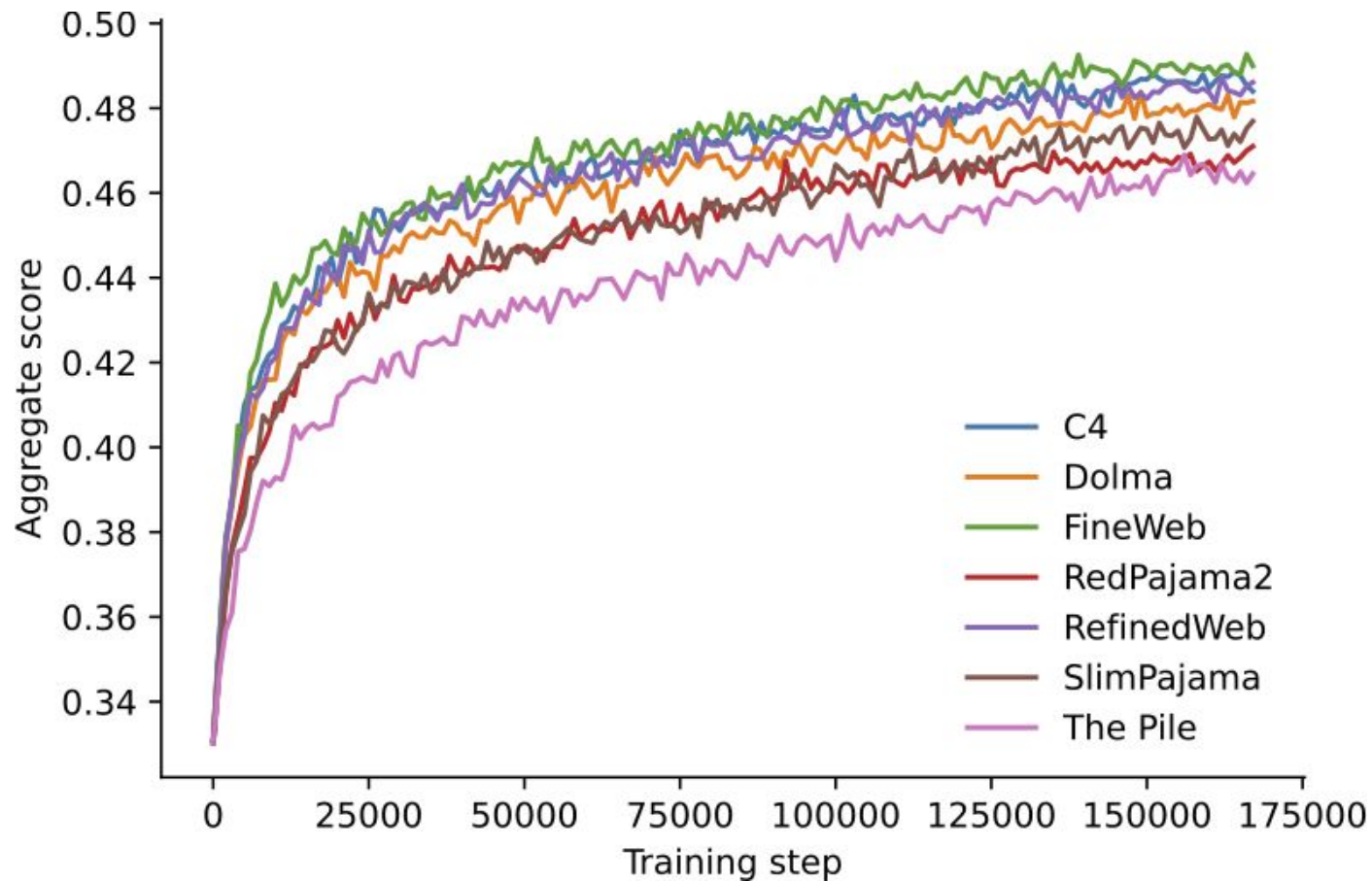
Full dataset statistics

- 25 sources
- 8TB of raw UTF-8 text
 - 60%, or 4.7TB of which is stackv2
- 4TB after filtering
 - stackv2 shrinks to 1.3TB or 33%
- 3.7TB after deduplication
 - 33% code, 25% patents, 20% scientific papers, 7% web, 15% "other"
- 1 trillion GPT-2 tokens

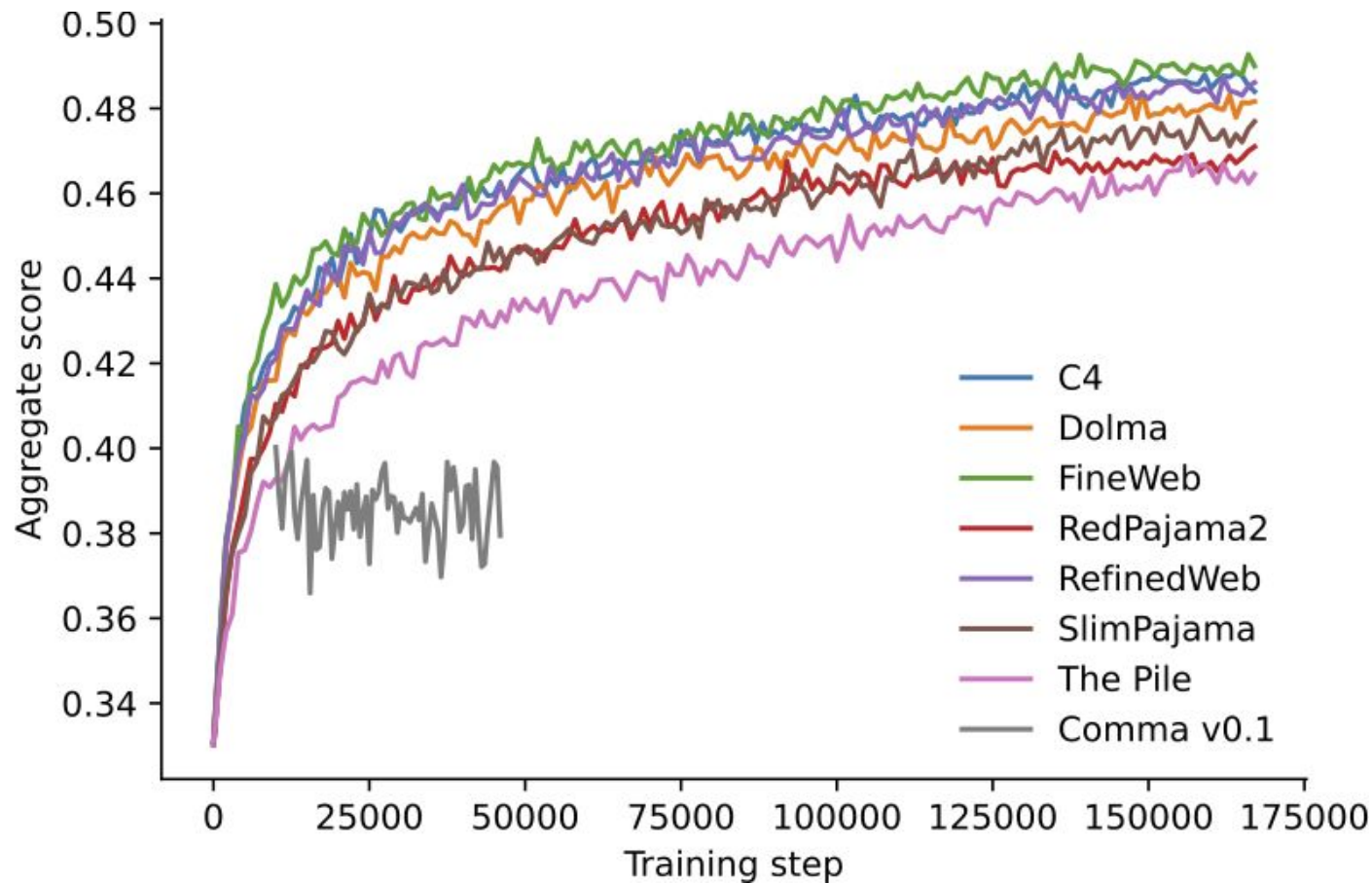
Initial Comma v0.1 ablation experiments

- Following the FineWeb ablation setup: 1.7B params, 350B tokens
- Standard Llama-style architecture
- Comparing to C4, Dolma (pre-1.6), FineWeb, RedPajama2, RefinedWeb, SlimPajama, The Pile
- Evaluating aggregate score on CommonsenseQA, HellaSWAG, OpenBookQA, PIQA, SIQA, Winogrande, ARC, and MMLU
- Initially simply concatenating all of the sources in the Common Pile

Ablation results (without Comma)



Ablation results (preliminary...)



Walking on a narrow bridge



Debugging by pondering

```
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1518, in _wrapped_call_impl
[ip-26-0-161-138:3]:     return self._call_impl(*args, **kwargs)
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1527, in _call_impl
[ip-26-0-161-138:3]:     return forward_call(*args, **kwargs)
[ip-26-0-161-138:3]: File "/fsx/craffel/nanotron/src/nanotron/nn/layer_norm.py", line 39, in forward
[ip-26-0-161-138:3]:     return layer_norm_fn(
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/flash_attn/ops/triton/layer_norm.py", line 875, in layer_norm_fn
[ip-26-0-161-138:3]:     return LayerNormFn.apply(
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/torch/autograd/function.py", line 539, in apply
[ip-26-0-161-138:3]:     return super().apply(*args, **kwargs) # type: ignore[misc]
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/flash_attn/ops/triton/layer_norm.py", line 748, in forward
[ip-26-0-161-138:3]:     y, y1, mean, rstd, residual_out, seeds, dropout_mask, dropout_mask1 = _layer_norm_fwd(
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/flash_attn/ops/triton/layer_norm.py", line 335, in _layer_norm_fwd
[ip-26-0-161-138:3]:     _layer_norm_fwd_1pass_kernel[(M,)](
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 100, in run
[ip-26-0-161-138:3]:     timings = {config: self._bench(*args, config=config, **kwargs)}
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 100, in <dictcomp>
[ip-26-0-161-138:3]:     timings = {config: self._bench(*args, config=config, **kwargs)}
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 83, in _bench
[ip-26-0-161-138:3]:     return do_bench(kernel_call, warmup=self.warmup, rep=self.rep, quantiles=(0.5, 0.2, 0.8))
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/testing.py", line 104, in do_bench
[ip-26-0-161-138:3]:     fn()
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 81, in kernel_call
[ip-26-0-161-138:3]:     self.fn.run(*args, num_warps=config.num_warps, num_stages=config.num_stages, **kwargs)
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 232, in run
[ip-26-0-161-138:3]:     return self.fn.run(*args, **kwargs)
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 232, in run
[ip-26-0-161-138:3]:     return self.fn.run(*args, **kwargs)
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/runtime/autotuner.py", line 232, in run
[ip-26-0-161-138:3]:     return self.fn.run(*args, **kwargs)
[ip-26-0-161-138:3]: File "<string>", line 65, in _layer_norm_fwd_1pass_kernel
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/compiler/compiler.py", line 579, in __getattr__
[ip-26-0-161-138:3]:     self._init_handles()
[ip-26-0-161-138:3]: File "/fsx/craffel/miniconda3/envs/exp/lib/python3.10/site-packages/triton/compiler/compiler.py", line 570, in _init_handles
[ip-26-0-161-138:3]:     mod, func, n_reqs, n_spills = fn_load_binary(self.metadata["name"], self.asm[bin_path], self.shared, device)
[ip-26-0-161-138:3]: RuntimeError: Triton Error [CUDA]: device-side assert triggered
[ip-26-0-161-138:2]: ../aten/src/Aten/native/cuda/Indexing.cu:1292: indexSelectLargeIndex: block: [26,0,0], thread: [96,0,0] Assertion `srcIndex < srcSelectDimSize` failed.
[ip-26-0-161-138:2]: ../aten/src/Aten/native/cuda/Indexing.cu:1292: indexSelectLargeIndex: block: [26,0,0], thread: [97,0,0] Assertion `srcIndex < srcSelectDimSize` failed.
```

Fixing by translating and pasting

既知の問題 (troubleshooting)

! 学習時に以下のようなエラーが発生します。どうすれば良いですか？

```
packages/triton/runtime/cache.py", line 78, in get_group
    raise Exception(f"Group file {p} does not exist from group {grp_filename}")
Exception: Group file /home/user/.triton/cache/8255a02245de2f538d4e742e52e5ee
```

triton cacheが複数のrankから参照されるとき、うまく参照ができず、発生するエラーです。
以下のように変更をすることで対処可能です。

まず、`.env/lib/python3.10/site-packages/triton/runtime/cache.py:L7` に以下を追加します。

```
+ import torch.distributed as torch_distributed
```

次に、`.env/lib/python3.10/site-packages/triton/runtime/cache.py:L91` を以下のようにします。

```
- temp_path = f"{filepath}.tmp.pid_{pid}_{rnd_id}"
- mode = "wb" if binary else "w"
- with open(temp_path, mode) as f:
-     f.write(data)
- # Replace is guaranteed to be atomic on POSIX systems if it succeeds
- # so filepath cannot see a partial write
- os.replace(temp_path, filepath)
+ # *** Rank 0 only ***
+ if torch_distributed.is_initialized() and torch_distributed.get_rank() == 0:
+     temp_path = f"{filepath}.tmp.pid_{pid}_{rnd_id}"
+     mode = "wb" if binary else "w"
+     with open(temp_path, mode) as f:
+         f.write(data)
```

Long debugging times



```
(exp) craffel@login-node-1:/fsx/craffel/comma-v0.1-ablations$ squeue --format="%18i %9P %30j %8u %8T %10M %9l %6D %R" --me
```







JOBID	PARTITION	NAME	USER	STATE	TIME	TIME_LIMI	NODES	ODELIST(REASON)
12171121_[0]	hopper-cp	merge-commonpile0p1nostacknodp	craffel	PENDING	0:00	2-02:00:00	1	(Dependency)
12171120_[0]	hopper-cp	merge-commonpile0p1nostacknodp	craffel	PENDING	0:00	2-02:00:00	1	(DependencyNeverSatisfied)
12171119_[0]	hopper-cp	tok-commonpile0p1nostacknodpi	craffel	PENDING	0:00	20:00:00	1	(DependencyNeverSatisfied)
12195297	hopper-pr	commav0p1all-6	craffel	PENDING	0:00	UNLIMITED	8	(Resources)

Caveats


- Data is too big to audit; can't be certain about licenses...
- Legality of nonpermissive pre-training text is still TBD
- Even if it's "legal" it might not be ethical
- Authors might not have given consent
- Still unclear how to provide attribution
- Some sources violate ToS
- Etc...




Collaborate/contribute!



r-three / common-pile







<> Code Issues 26 Pull requests 8 Discussions Actions Projects ...






common-pile Public Edit Pins Unwatch 6 Fork 6 Star 21

main Go to file + <> Code About 

**blester125** Add Dolma Counter scri... ... ✓ de0dad9 · last month 

 .github/workflows	add ci for linting	3 months ago
 arxiv	Create README.md	6 months ago
 bhl	run linters over all files	3 months ago
 courtlistener	USPTO (#71)	last month

Repo to hold code and track issues for the collection of permissively licensed data

-  Readme
-  MIT license
-  Activity
-  Custom properties
-  21 stars

