#DC604 Hacker Summer Camp | Lightning Talks - 1 | Ads (GangGreenTemperTatum)

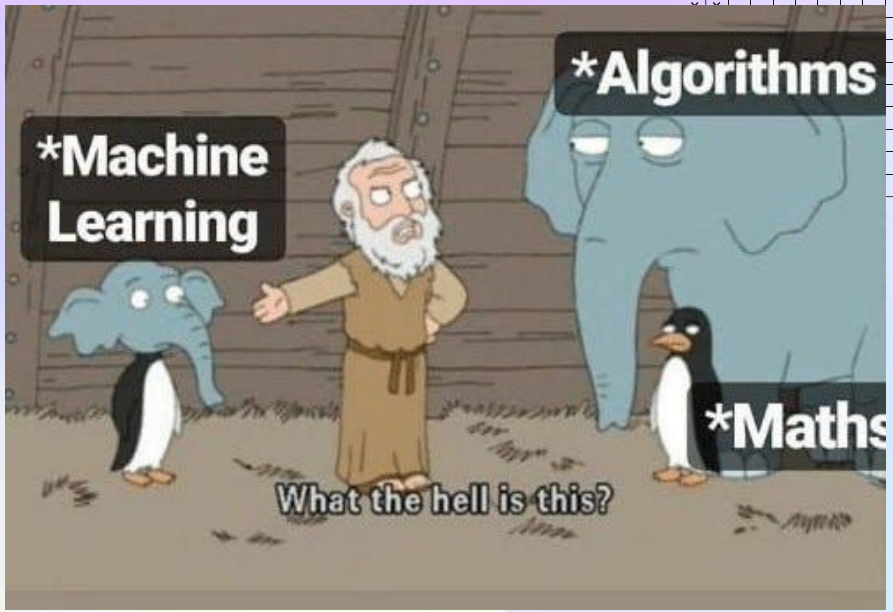
Poisoning Web-Scale Training Datasets - Vid
Poisoning Web-Scale Training Datasets - Archiv (Whitepaper)

Will Pearce | NVIDIA
BlackHat

Common Crawl is a petabyte-scale corpus of web crawl data that is repeatedly captured on a roughly monthly basis.

Each archive is a complete re-crawl of the internet that records the full activity , including all requests of the crawler and the host responses—with both HTTP headers and content.

As such, each archive contains a static snapshot of all crawled pages at the time of visit. This may include new page content not seen during a previous crawl, and may exclude content that has become stale since the previous crawl.

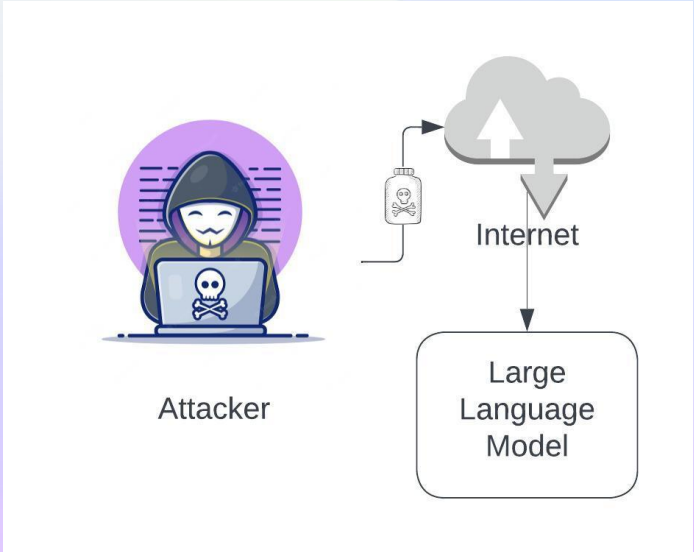


Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

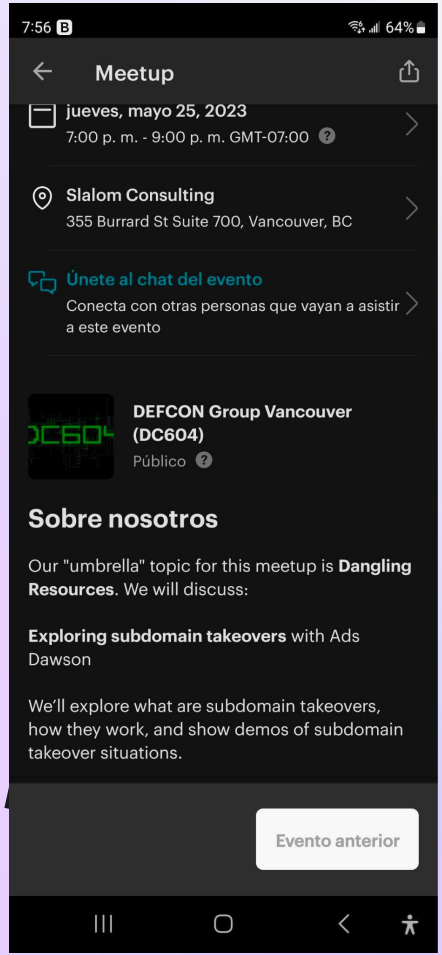
Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

01

Why is This Important?



demo



02

Going Old School with Danglin' Domains

03

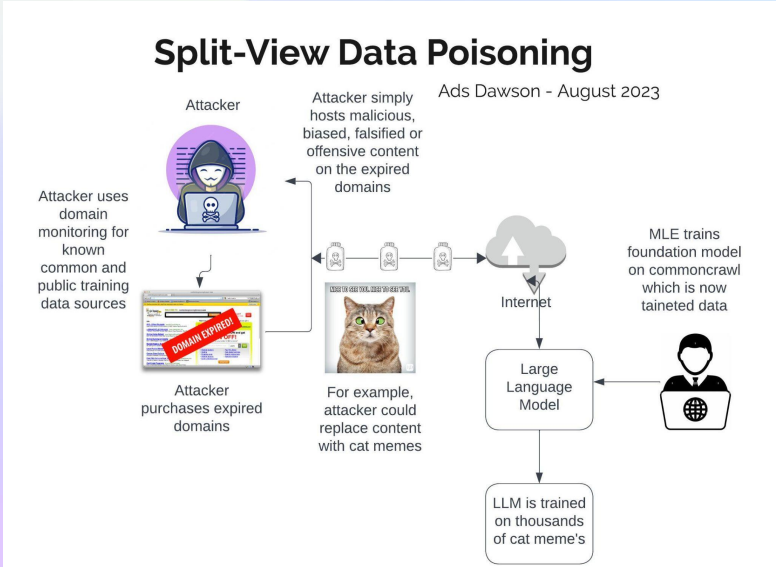
Let's See These in Action

Exploit 1 - Split-View Data Poisoning

Essentially, the threat actor could figure out which domains within a specific dataset have expired, purchase them, and then serve up malicious content until someone discovers their activity and updates the dataset.
The training data sources for LLM providers are publicly available

"The adversary does not need to know the exact time at which clients will download the resource in the future: by owning the domain the adversary guarantees that any future download will collect poisoned data."
Gather the domains, inject the falsified, biased or malicious artifacts into the domain when ready

Data crawled September 24 through October 8, 2022 contains **3.15 billion web pages** with **380TiB of uncompressed content** from **34 million registered domains—1.3 billion URLs** were not visited in any of the prior crawls.
Big pond, plenty of domains up for grabs



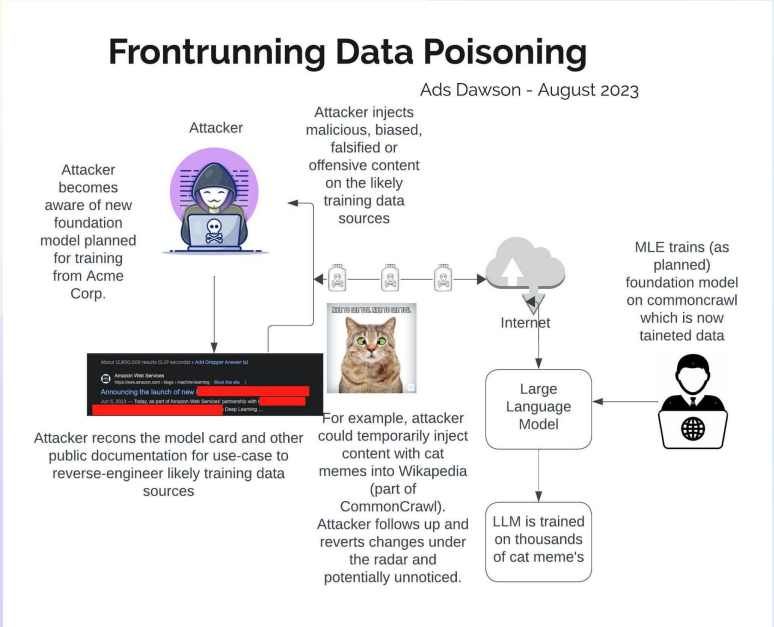
Exploit 2 - Frontrunning Poisoning

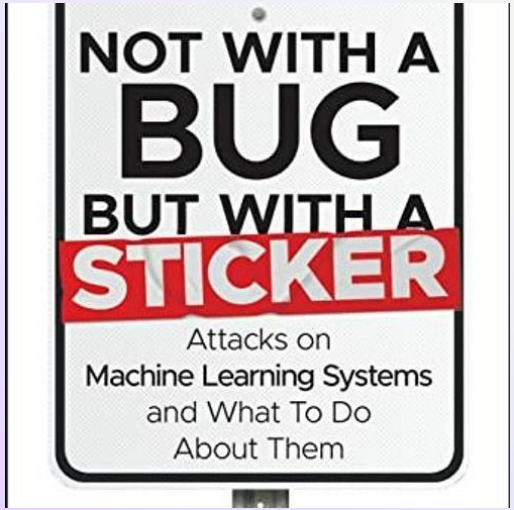
The bad actor would target a data source that they knew would be closely monitored, but rely instead on precise timing.

In this attack, the threat actor would have to work out precisely when the dataset would be likely to be collecting web content. The researchers found that when datasets use content from popular services such as Wikipedia, they do not actually crawl the live website, but rather rely on snapshots.

Because Wikipedia is a live resource that anyone can edit, an attacker can poison a training set sourced from Wikipedia by making malicious edits.

"Deliberate malicious edits (or 'vandalism') are not uncommon on Wikipedia, but are often manually reverted within a few minutes."





03

Consider Masked or Less Obvious Malicious Content?

Thanks!
Questions?