"A robot walks into a library..."

Common Crawl would like to submit a response to the recent inquiry from the US Copyright Office regarding AI training.

Our goal is to protect the **Right to Read,** on behalf of people, and machines. It is a pivotal time for the development of LLM technology.

Robots designed to self-improve should be granted **Access to Knowledge**. Providing the broadest training data is the best way to ensure the creation of agents capable of better understanding and benefiting human needs.

It is imperative that we establish a framework ensuring **Equitable Access to Data**, acknowledging the crucial role that crawled data plays in machine learning and human development. Policies should be carefully constructed, and should strike a balance between copyright protection and AI advancement.

To hinder the development of intelligent machines ultimately hinders the development of humanity.

Embracing the use of copyrighted materials in AI training data is a pragmatic approach in a global context – by adopting a balanced approach to copyright, we can maintain a competitive edge in AI research and development, and set a global example.

While it is important to respect intellectual property rights, implementing strict regulations without due consideration may disadvantage our technological leadership.

- Common Crawl is the Primary Training Dataset for every LLM
- 82% of raw tokens used to train GPT-3
- 240 billion web pages spanning 16 years
- 5 billion new pages added each month
- Founded in 2007 by Gil Elbaz, 501(c)(3) nonprofit

- 10,000 academic citations dating back to 2008
- Interview with Gil Elbaz from 2012
- Common Crawl blog posts
- CCBot - Common Crawl's web crawler
- Sebastian's Feb 2023 talk about Common Crawl
- NYU student paper from 2013

- Common Crawl is a National Treasure
- Common Crawl facilitates important technical advances for us - and everyone
- Common Crawl is a modern Library of Alexandria - it is the record of our civilization

Common Crawl archives documents that have been published on the open public Internet. It crawls politely - obeying the Robots Exclusion Protocol (robots.txt), a standard established in 1994. It does not bypass paywalls, log into websites, or accept cookies. It only crawls documents that have been intentionally published on the web in order for people to be able to see them. CCbot identifies itself and goes away if robots.txt tells it to. Note the comments in our technical talk about the lengths CCbot goes to be a good citizen of the web and a respectful web crawler.

Threads from news.ycombinator.com:

Ask HN: What would be the fastest way to grep Common Crawl? - https://news.ycombinator.com/item?id=22214474 - Feb 2020 (7 comments

Using Common Crawl to play Family Feud - https://news.ycombinator.com/item?id=16543851 - March 2018 (4 comments)

Web image size prediction for efficient focused image crawling - https://news.ycombinator.com/item?id=10107819 - Aug 2015 (5 comments)

102TB of New Crawl Data Available - https://news.ycombinator.com/item?id=6811754 - Nov 2013 (37 comments)

SwiftKey's Head Data Scientist on the Value of Common Crawl's Open Data [video] - https://news.ycombinator.com/item?id=6214874 - Aug 2013 (2 comments)

A Look Inside Our 210TB 2012 Web Corpus - https://news.ycombinator.com/item?id=6208603 - Aug 2013 (36 comments)

Blekko donates search data to Common Crawl - https://news.ycombinator.com/item?id=4933149 - Dec 2012 (36 comments)

Common Crawl - https://news.ycombinator.com/item?id=3690974 - March 2012 (5 comments)

CommonCrawl: an open repository of web crawl data that is universally accessible - https://news.ycombinator.com/item?id=3346125 - Dec 2011 (8 comments)

Tokenising the english text of 30TB common crawl - https://news.ycombinator.com/item?id=3342543 - Dec 2011 (7 comments)

Free 5 Billion Page Web Index Now Available from Common Crawl Foundation - https://news.ycombinator.com/item?id=3209690 - Nov 2011 (39 comments)

Background:

- US Copyright Office wants to hear what people think about AI and copyright
- Foundation Models and Fair Use (Stanford paper)
- Prof Samuelson of Berkeley Large Language Models Meet Copyright Law

–
Rich Skrenta
Executive Director, Common Crawl Foundation