NO PREP

The Internet as a Textbook

Where AI Gets Its Information

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

The Internet as a Textbook: Where AI Gets Its Information

The World's Largest, Messiest Library

Imagine a library with billions of books, but no librarian, no catalog system, and no quality control. Anyone can add anything. Some of the books are brilliant academic research. Some are conspiracy theories. Some are advertisements written to look like news articles. Some are plagiarized copies of other books. This is a reasonable description of the internet, and it is the primary source of training data for most large language models. One of the most important datasets used to train LLMs is called Common Crawl. It is a nonprofit organization that has been systematically downloading and archiving web pages since 2008. Common Crawl

Name: **Period:** **Date:**

The Internet as a Textbook: Where AI Gets Its Information

The World's Largest, Messiest Library

Imagine a library with billions of books, but no librarian, no catalog system, and no quality control. Anyone can add anything. Some of the books are brilliant academic research. Some are conspiracy theories. Some are advertisements written to look like news articles. Some are plagiarized copies of other books. This is a reasonable description of the internet, and it is the primary source of training data for most large language models. One of the most important datasets used to train LLMs is called Common Crawl. It is a nonprofit organization that has been systematically downloading and archiving web pages since 2008. Common Crawl releases its data publicly, and it contains petabytes of raw web content: hundreds of billions of web pages in dozens of languages. AI companies use Common Crawl as a foundation for training data because it is enormous and freely available.

What Gets Included (and What Gets Cut)

Raw web data is far too messy to use directly. Before training, AI companies apply filters to remove low-quality content. Common techniques include removing duplicate pages, filtering out content that appears to be spam or auto-generated garbage, removing pages with very little text, and excluding sites known for harmful or illegal content. Even after filtering, what remains is a reflection of the internet itself: content written primarily in English, produced disproportionately by people in wealthy countries, skewed toward topics that are popular online, and filled with the biases, assumptions, and perspectives of whoever wrote it. Less-represented languages, cultures, and viewpoints get fewer pages on the internet and therefore less representation in the training data. This is a fundamental problem with using the internet as a textbook.

Books and Academic Sources

Beyond the web, AI companies have trained on digitized books. Projects like Project Gutenberg have digitized millions of books with expired copyrights, making them freely available. OpenAI reportedly used two book datasets, called Books1 and Books2, in training earlier versions of GPT. However, reports have suggested that some of these datasets may have included copyrighted books without permission from the authors, which has led to multiple ongoing lawsuits from writers and publishers. Academic papers represent another significant source. Papers published on preprint servers like arXiv, which publishes millions of open-access papers in physics, math, computer science, and other fields, have been widely used in AI training. Wikipedia is also a standard component because it offers well-organized, factual text on millions of topics across many languages. However, Wikipedia itself has known gaps, biases, and quality variations across different topics and languages.

Code as Training Data

Modern LLMs are not just trained on human language. Many have been trained on enormous repositories of programming code. GitHub, the world's largest code-sharing platform, contains hundreds of millions of public code repositories written by programmers around the world. This code was used to train models like GitHub Copilot, which can generate programming code in response to natural language descriptions. Training on code gives LLMs an interesting capability: they can reason about structured logic and follow rule-based patterns more reliably than if they had been trained on text alone. It is also why modern LLMs perform surprisingly well at math problems, spreadsheet formulas, and logical reasoning tasks. The patterns of formal structure in code seem to reinforce the model's ability to handle structured thinking in other domains.

The Copyright Problem

The internet contains an enormous amount of copyrighted content: news articles, song lyrics, book excerpts, movie scripts, and more. When AI companies scraped this content for training data, they entered legally murky territory. Copyright law in the United States gives creators exclusive rights to their work. Whether training an AI model on copyrighted text counts as infringement is currently being decided in courts around the world. The New York Times sued OpenAI and Microsoft in 2023,

arguing that its copyrighted articles were used to train ChatGPT without permission or compensation. The Authors Guild has filed lawsuits on behalf of writers. Visual artists have sued AI image generators. These cases are still working through the courts, and their outcomes will significantly shape how AI companies gather training data in the future.

What Training Data Means for Users

Understanding where AI training data comes from matters for anyone who uses these tools. If you ask an LLM about a topic that is underrepresented on the internet, such as indigenous languages, local history, or non-English cultural practices, the model may give you thin or inaccurate responses because it had less training data on those topics. If you ask about something that has lots of misleading content on the internet, the model may reflect that misinformation. The training data also determines what the model knows as normal. Because English-language content dominates the web, LLMs often perform better in English than in other languages. Because certain perspectives dominate online discourse, those perspectives may feel more 'natural' to the model. Recognizing these built-in limitations helps you use AI tools more critically and understand why their outputs should always be verified against diverse, authoritative sources. Multilingual challenges in AI training reveal just how unequal the internet's language distribution really is. According to estimates, roughly 55 percent of all internet content is in English, despite English being the native language of less than 6 percent of the world's population. French, German, Russian, and Spanish are significantly represented but still far below English. Languages like Swahili, Bengali, and thousands of smaller languages are represented by only a tiny fraction of available web content. This means AI models trained on internet data encode a fundamentally skewed view of human knowledge and culture, one that reflects the perspectives, values, and priorities of the populations most active in English-language online spaces. The consent problem in AI training data extends well beyond copyright. Many people whose personal writing, social media posts, comments, or emails ended up in training datasets never had any awareness that their text would be used to train commercial AI systems. In Europe, the GDPR's principles of data minimization and purpose limitation raise questions about whether scraping public internet content for AI training purposes constitutes lawful processing of personal data.

Several European data protection authorities have investigated AI companies on these grounds. In the United States, where data protection law is more fragmented, the legal landscape is less clear but evolving rapidly as lawmakers and regulators pay increasing attention to how AI training data is gathered. Understanding training data sources also has implications for privacy. Much of the personal information that people share online, from forum posts to social media comments to product reviews, may have ended up in training datasets without those individuals ever realizing it. When an AI model produces text that feels eerily specific to a certain personal situation or mirrors language patterns from a particular online community, it may be because the model encountered similar text during training. Privacy advocates argue that people should have the right to know whether their public online writing was used to train commercial AI systems, and in some jurisdictions this right is beginning to be recognized in law.

Key Information: AI Training Data Sources

Detail	Information
Primary web dataset	Common Crawl: petabytes of archived web pages since 2008
Book sources	Project Gutenberg (public domain), Books1/Books2 datasets
Academic sources	arXiv (open-access papers), Wikipedia
Code source	GitHub: hundreds of millions of public code repositories
Data filtering goal	Remove spam, duplicates, low-quality, and harmful content
Language bias	English-language content vastly overrepresented
Key legal issue	Copyright lawsuits from publishers, authors, and news organizations
NYT lawsuit	New York Times sued OpenAI and Microsoft in 2023

Detail	Information
Effect on underrepresented topics	Less training data = thinner, less accurate model outputs
Data curation	Process of selecting and cleaning training data to improve quality

Case Study: Amara's Research on Her Grandmother's Language

Amara is working on a heritage project for her social studies class. Her grandmother speaks Tigrinya, a language spoken by several million people in Eritrea and Ethiopia. For her project, Amara decided to use an AI chatbot to help her gather information about Tigrinya proverbs, traditions, and historical background. The AI struggled. It provided a few general facts about Eritrea but admitted it had limited information on Tigrinya-specific cultural traditions. When Amara asked it to translate a phrase into Tigrinya, the output looked wrong to her grandmother, who said the translation was awkward and partly incorrect. Amara was frustrated. She had seen her classmates use the same AI to easily research European and American history and get detailed, polished results. She asked her teacher why the AI worked so much better for some topics than others. Her teacher explained that the AI was trained primarily on internet content, and Tigrinya-language content on the internet is extremely limited compared to English. The AI simply had far less data to learn from.

1. Using what you learned about training data sources, explain why the AI performed poorly when Amara asked about Tigrinya language and culture.

2. What does Amara's experience reveal about the limitations of using AI as a research tool for topics outside mainstream Western culture?

Multiple Choice

1. What is Common Crawl?

    a) A major social media platform used to gather public opinions for training data

    b) A Google product that indexes and ranks websites for search

    c) A nonprofit that archives billions of web pages used in AI training datasets

    d) A software tool that filters harmful content from AI outputs

2. Why did major media organizations like The New York Times sue AI companies in 2023?

    a) Because AI companies refused to buy significant advertising space in traditional newspapers

    b) Because AI companies used their copyrighted articles as training data without permission

    c) Because AI chatbots were giving away free news content

    d) Because AI chatbots were outperforming human journalists

3. Why do LLMs generally perform better in English than in other languages?

    a) Non-English speakers around the world rarely use or have reliable access to AI tools

    b) English is programmed into LLMs as the default language

    c) AI companies deliberately exclude non-English content to reduce complexity

    d) English-language content vastly dominates the internet, providing more training data

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Common Crawl	A) The process of selecting, cleaning, and filtering training data to improve quality
2. ___ Data curation	B) A nonprofit organization that has archived billions of web pages used in AI training
3. ___ arXiv	C) Legal protection giving creators exclusive rights to their work
4. ___ Copyright	D) The world's largest code-sharing platform, a major source of programming training data
5. ___ GitHub	E) An open-access repository of academic papers widely used in AI training

Fill-in-the-Blank

1. The principle that the internet contains far more English content than other languages creates a _____ problem in AI training data.

2. The New York Times filed a _____ against OpenAI in 2023 over use of its articles as training data.

3. Training on _____ repositories like GitHub helps LLMs reason about structured logic and formal patterns.

Short Answer

Answer each question in 2-4 complete sentences.

1. Why is the raw data from Common Crawl not used directly for training, and what steps do AI companies take to improve it?

2. How does training on programming code improve an LLM's performance on non-coding tasks like math and logical reasoning?

3. If you were an AI company designing a more equitable training dataset, what kinds of changes would you make to the current approach?