



Acquiring High- Quality Web Data for LLM Fine-Tuning

Introduction

In the realm of AI, Large Language Models (LLMs) stand as titans, reshaping how machines understand and generate human language.

If you're an AI specialist, you're already familiar with LLM capabilities and the huge impact they had over the last few years. You must also know that perfecting the model is all about the fine-tuning process. For that, you need access to vast amounts of high-quality data.

With the amount of data required for LLM fine-tuning, the data acquisition process gets even more difficult than usual. How do you ensure that the data is fresh, relevant, and of high quality? What if the process of acquiring it gets interrupted?

In today's whitepaper, we'll be answering all these questions and more. We'll begin by defining some common terms, then go over different data sources used for LLM fine-tuning. Finally, we'll explore the different issues AI specialists may run into while acquiring data for LLMs and offer solutions to those very issues.

Buckle in and prepare to dive deep into the complexities of data management for LLM optimization and discover effective strategies to overcome these challenges.

Table of contents

Fine-tuning vs. training vs. RAG	4
Data sources for fine-tuning LLMs	5
1. Licensed data	5
2. Data covered by public copyright licenses	6
3. Publicly available data	7
Acquiring and structuring web data for LLMS: challenges and solutions	8
1. Data volume	8
2. Anti-scraping measures	10
3. High-costs	11
4. Dynamic content, JavaScript rendering, different HTML structures	12
5. Unstructured or low-quality data	13
6. Legal considerations	14
The future of LLM fine-tuning: prioritizing data quality, privacy, and ethical AI practices	15
Final word	16

Fine-tuning vs. training vs. RAG

When working with LLMs, you'll likely encounter terms like training, fine-tuning, and the innovative RAG (Retrieval-Augmented Generation) method.

Training lays the groundwork, fine-tuning hones the model for specific tasks, and RAG enhances its capabilities by incorporating real-time information.

Training is the beginning process where a general language model is built using vast, diverse data to understand language broadly. Meanwhile, **fine-tuning** takes that pre-trained model and adapts it for specific tasks or domains using a smaller, targeted dataset, making it more specialized and effective in those areas.

It's also important to note that fine-tuning doesn't rely on continuous data collection; rather, it uses **periodic data snapshots**. This approach provides a stable dataset to refine the model, with adjustments only made when a new snapshot is taken to reflect updated information. Since fine-tuning is time-intensive, by using these snapshots you have better control over the model's optimization process without requiring constant updates.

As of recently, **RAG (Retrieval-Augmented Generation)** has emerged as a new method for enhancing LLMs by retrieving relevant external data in real-time during text generation. It ensures more accurate, up-to-date responses while leveraging the model's general knowledge.

In today's whitepaper, we're mostly going to talk about the process of fine-tuning your model.

Data sources for fine-tuning LLMs



Whether you're in the stage of training or fine-tuning your LLM, you'll need vast and diverse data. This way, you're helping the model capture nuanced, domain-specific knowledge. The quality and variety of these data sources directly impact the model's ability to perform well in specific applications. Let's take a look at some primary data sources (and their sub-categories) used for training or fine-tuning LLMs:

1. Licensed data

This type of data is protected by copyright or other intellectual property rights and can only be used with permission from the owner. The owner grants specific rights to use the data through a license agreement that outlines the terms and conditions of use. Licensed data may have restrictions on commercial use, redistribution, modification, etc.

2. Data covered by public copyright licenses

Data covered by public copyright licenses is data that's available for anyone to access, use, modify, and share for purposes defined in the license under specific license terms. This data is usually licensed under a public copyright license, such as Creative Commons or Open Data Commons licenses, which grant you broad usage rights with minimal restrictions. The goal of such data is to enable collaboration, transparency, and innovation. Let's explore a few key examples of such data contributors to LLM development:

- **Wikipedia:** The English Wikipedia is often used as a foundational dataset for fine-tuning LLMs thanks to its endless amount of structured and unstructured information. The data is available under a [Creative Commons license](#), so it can be used for research and development.
- **Common Crawl:** This is a massive repository of web data collected through web crawling. It provides a diverse set of text from the internet, which can be used for fine-tuning LLMs. The data is freely available and can be processed to suit specific model requirements.
- **Hugging Face:** A community-driven platform offering a wide range of ready-to-use datasets for fine-tuning LLMs. It supports multiple data types (text, images, audio) and integrates with ML frameworks like PyTorch and TensorFlow. The library simplifies dataset loading, preprocessing, and management. It's a valuable resource for efficient data preparation that's later used for LLM-fine tuning.
- **The Pile:** Developed by EleutherAI, The Pile is a large-scale dataset specifically designed for training and fine-tuning LLMs. With over 800 GB of data from 22 diverse sources (which include academic papers, web pages, books, and code) it offers a broad, balanced dataset that enhances model performance across various tasks. The Pile's diversity and scale makes it an excellent resource for fine-tuning LLMs, ideal for adapting models to specialized applications.

Such resources not only democratize access to vast information but also empower researchers to accelerate breakthroughs in AI performance and capability.

3. Publicly available data

Publicly available data* term refers to the information that's freely accessible on the internet, visible to any user without requiring login credentials, subscriptions, or paywalls.

Distinction from data available under public copyright license

While publicly available data is available to the public, it is important to note that not all publicly available data is considered as data licenses under public copyright license. Data available under public copyright license is a subset of publicly available data that is specifically structured and formatted to be easily accessible and usable by anyone, often under open licenses that encourage sharing and collaboration.

In summary, publicly available data is a broad category of information made available for public use, promoting transparency and informed decision-making, while **data available under public copyright license** represents a more refined and accessible subset of this information.

For example...

- **E-commerce product listings:** Public product descriptions, reviews, and pricing information from marketplaces like Amazon, eBay, and public Shopify stores.
- **News websites and articles:** Publicly accessible articles and reports from media outlets (e.g., BBC, Reuters, CNN) that cover various topics such as politics, technology, and global events.
- **WHO data:** WHO data is a type of data that falls under both data available under public copyright license and publicly available data. It qualifies as both publicly available and data available under public copyright license data due to its availability to the general public without restrictions and its structured format that facilitates easy access and use. This dual classification supports the goals of transparency, public health improvement, and collaborative research efforts.

*Disclaimer: It's important to note that while publicly available data is often accessible, it can still be subject to licensing agreements or terms of service that govern its use. Scraping such data without proper consideration can result in legal complications. Always ensure compliance with the relevant site policies and licensing terms before engaging in large-scale data collection, and consult legal advice when necessary to avoid unintended violations.

Acquiring and structuring web data for LLMs: challenges and solutions

Now, let's get into the meat and potatoes of this whitepaper: what makes data acquisition for LLM fine-tuning so difficult? And how do you deal with that? There are several challenges you'll run into while acquiring data, some generic (CAPTCHAs, for example), some exclusive to LLM fine-tuning use case specifically.



1. Data volume

One of the most significant challenges in fine-tuning LLMs is acquiring the sheer volume of data required to improve model accuracy. LLMs thrive on vast datasets, but gathering such enormous quantities of web data presents technical and logistical hurdles.

Efficient web scraping at this scale requires advanced techniques capable of handling immense data loads without overwhelming server resources or triggering rate limits. Additionally, scraping systems must be scalable, ensuring that data collection processes can be expanded as needed to accommodate the growing demands of LLM fine-tuning.

Real-life examples: OpenAI and Google Chinchilla

To look at real-life examples, **OpenAI's GPT-3** alone was trained on around 570GB of curated text, representing roughly 300 billion tokens. To put this in perspective, that's about as much data as 500 hours of HD video streaming—equivalent to **watching non-stop for 21 days**—or the text of **570,000 books, which would take a lifetime to read**.

Similarly, a trillion-token model, like **Google's Chinchilla**, requires even more extensive datasets, comparable to the entire text of **1.4 million books**. If each page were printed and stacked, the resulting pile would tower more than 30 kilometers (or 19 miles) high, dwarfing even Mount Everest.



Solution: Datacenter Proxies

Generally speaking, all Oxylabs scraping solutions can handle large scale scraping. It's more so the matter of how sensitive the target website is.

Let's take our [Datacenter Proxies](#) for example: starting at \$0.65 /GB, these proxies make an ideal budget-friendly solution for large scale scraping.

Here's why:

- All our scraping infrastructures (Datacenter Proxies included) are robust and highly efficient. They're also monitored 24/7 by our professionals for any service disruptions.
- Our Datacenter Proxies are pre-tested; some are precisely selected for scraping specific targets.
- Unlike other proxy providers, our proxies have unlimited connections and concurrency.
- With the pay-per-IP billing model, our Datacenter Proxies offer entirely unlimited traffic and bandwidth.
- We have flexible agreement terms, offer custom plans, and are able to meet the client's required number of Datacenter IPs (100K and more.)



2. Anti-scraping measures

CAPTCHAs are designed to differentiate between human users and automated scripts. Advanced anti-bot mechanisms can detect scraping activities based on patterns such as request frequency, user agent, and interaction behaviors. These mechanisms can severely limit or block data acquisition processes.

Also, websites often employ IP blocking or rate limiting to prevent excessive requests from the same IP address, which is a common indicator of automated scraping. Rate limiting restricts the number of requests that can be made in a given time frame.



Solution: Web Unblocker, proxies, Web Scraper API

- [Web Unblocker](#) allows you to appear as a real user with the help of dynamic fingerprinting technology. Our system will select the right headers, cookies, JavaScript rendering, and other browser parameters for you.
- All of our scraping solutions feature automatically rotating IP addresses and ML-driven proxy selection. IP rotation helps prevent getting blocked during web scraping by changing the IP address used for each request. This technique distributes requests across multiple IP addresses, reducing the likelihood of exceeding rate limits or triggering anti-scraping mechanisms. You should be able to prevent CAPTCHA and, in turn, not slow down the response time (with CAPTCHA, the response time gets longer by 40-50%).
- Certain targets are highly sensitive and may trigger anti-scraping mechanisms anyway. To prevent it and to not waste your budget, you may need to use [Residential Proxies](#). By imitating human-like behavior, these proxies allow users to tackle even the most sensitive targets.



3. High-costs

Let's look back at the first point – the sheer volume of data required for LLM fine-tuning means high costs. Collecting, processing, and storing large datasets often requires a lot of money, especially if you're aiming for high-quality, up-to-date information. Not to mention, the extra costs of maintaining your web scraping infrastructure and the power necessary to process and structure these datasets. Balancing data quality and cost-efficiency becomes crucial to ensure a sustainable, scalable approach to acquiring data.



Solution: several web scraping solutions combined

- Like we mentioned in the last section, to tackle certain targets, you may require Residential Proxies specifically.
- On the other hand, many targets can be accessed through Datacenter Proxies, a budget-friendly solution.
- Our technical professionals can set you up an infrastructure with different products combined. This way, you can have an uninterrupted data stream for fine-tuning your LLM, spending a reasonable, optimal amount of money.
- Like we already mentioned, our scraping infrastructures are constantly monitored by our professionals – you won't need to worry about any additional costs or efforts to maintain a proper data stream.



4. Dynamic content, JavaScript rendering, different HTML structures

A significant challenge in web data acquisition for LLM fine-tuning is dealing with dynamic content and JavaScript-heavy websites. Many modern websites use JavaScript to load or update content dynamically, meaning the data needed for fine-tuning might not be visible in the initial HTML response.

Extracting this content requires sophisticated techniques to render JavaScript and access the data behind it. Additionally, different websites can have varied HTML structures, making it difficult to apply uniform scraping rules across sites.



Solution: OxyCopilot

- [OxyCopilot](#) is the first-of-a-kind AI web scraper assistant designed to generate scraping codes and parsing templates using simple prompts in English. It leverages Oxylabs' [Web Scraper API](#) and saves developers hours of manual coding by utilizing advanced LLMs to analyze complex HTML structures, find the best parsing patterns, and generate parsing instructions and API request codes.
- OxyCopilot easily adapts to changes in website structures and returns rendered data. With other tools, you may need to wait for days for the code to be adjusted to a new website structure. Meanwhile, OxyCopilot's code is automatically updated, so your scraping projects are not disrupted.



5. Unstructured or low-quality data

One of the more persistent challenges in acquiring web data for LLM fine-tuning is unstructured or low-quality data. Much of the information available on the web is not presented in a clean, standardized format, often appearing as free-form text, noisy data, or inconsistent structures. This can include duplicated content, irrelevant information, incomplete records, or poor formatting, all of which complicate the data extraction and processing pipeline.

For the fine-tuning process, it's essential that the model learns from meaningful, accurate information. To achieve this, you may have to put a lot of effort into cleaning and preprocessing tasks, such as filtering out noise, removing duplicates, normalizing formats, and addressing inconsistencies.



Solution: OxyCopilot

- By default, [OxyCopilot](#) is designed to deliver well-structured, high-quality data (unless you specifically request otherwise.)
- Generally, our tools deliver the exact data as indicated in your request – no more, no less. If you're having trouble scraping and you're not receiving what you need, our [support team](#) is there to help 24/7 via live chat.



6. Legal considerations

Finally, there's a lot of discussion about legal and ethical matters in the context of LLMs and AI in general; especially when it comes to personal or sensitive data. According to regulations like GDPR, you have to follow strict requirements on handling data that could identify individuals, including rules on consent, data minimization, and user privacy. Violating these regulations can result in significant penalties and damage to an organization's reputation.

Our approach:

- To ensure the client's use case is legit, for larger customers, we do standard KYC (Know Your Customer) procedures.
- We have established a clear Agreement that outlines the acceptable use policy for our services.
- We have a list of restricted targets that include governmental, financial data, and so on. More information [here](#).
- We have implemented rigorous risk assessments and policies aligned with the ISO/IEC 27001:2017 standard for Information Security Management Systems to uphold data integrity and security. All processes are designed to ensure adherence to this standard, reinforcing our strong commitment to best practices in information security.
- The selection of data to scrape is determined by the client's specific needs and objectives. To ensure that all scraping activities align with relevant legal and regulatory requirements, we strongly recommend that clients seek the advice of a legal professional prior to proceeding.

The future of LLM fine-tuning: prioritizing data quality, privacy, and ethical AI practices

In preparation for this white paper, we collaborated with Mantas L., AI Tech Lead. When asked for future predictions in the LLM landscape, here's what he had to say:



As AI models advance, prioritizing high-quality, contextually relevant data becomes critical. Instead of focusing on volume, organizations are innovating with AI-driven web scraping tools and adaptive algorithms to handle diverse web content more effectively. With rising global data privacy concerns, stricter regulations are likely, pushing companies to adopt privacy-preserving techniques that maintain data integrity while ensuring compliance.

We foresee a shift towards multi-modal data integration, merging text, audio, video, and image data, which demands sophisticated data-handling capabilities. Automation in data curation and cleaning will grow, yet ethical obligations will require organizations to ensure transparency and accountability in their data usage.

Real-time data processing is becoming essential for applications like customer service and analytics, necessitating timely responses.

However, the increase in AI-generated content risks diminishing the originality of training data, raising concerns about bias and the Echo Chamber Effect, where models trained on AI-generated data may lack diversity.

To counter these challenges, leveraging diverse and high-impact datasets will be crucial for developing robust, adaptable models. This comprehensive approach is vital for navigating AI's evolving landscape and ensuring ethical and innovative advancements in data management.

Final word

Acquiring high-quality web data for fine-tuning LLMs is challenging, involving large data volumes, dynamic content, and legal considerations. However, with a custom scraping infrastructure that works best for your priorities and targets, it's possible to train your language model effectively and in a cost-efficient manner. We hope that this whitepaper has shed some light on how Oxylabs web scraping solutions can help with common LLM training challenges.



Want to Know More?

If you are an AI specialist looking for solid data acquisition solutions, reach out to us at sales@oxylabs.io (or click a button below). We'll give you a custom offer that best fits your goals and target websites.

 **Get in touch with Oxylabs**