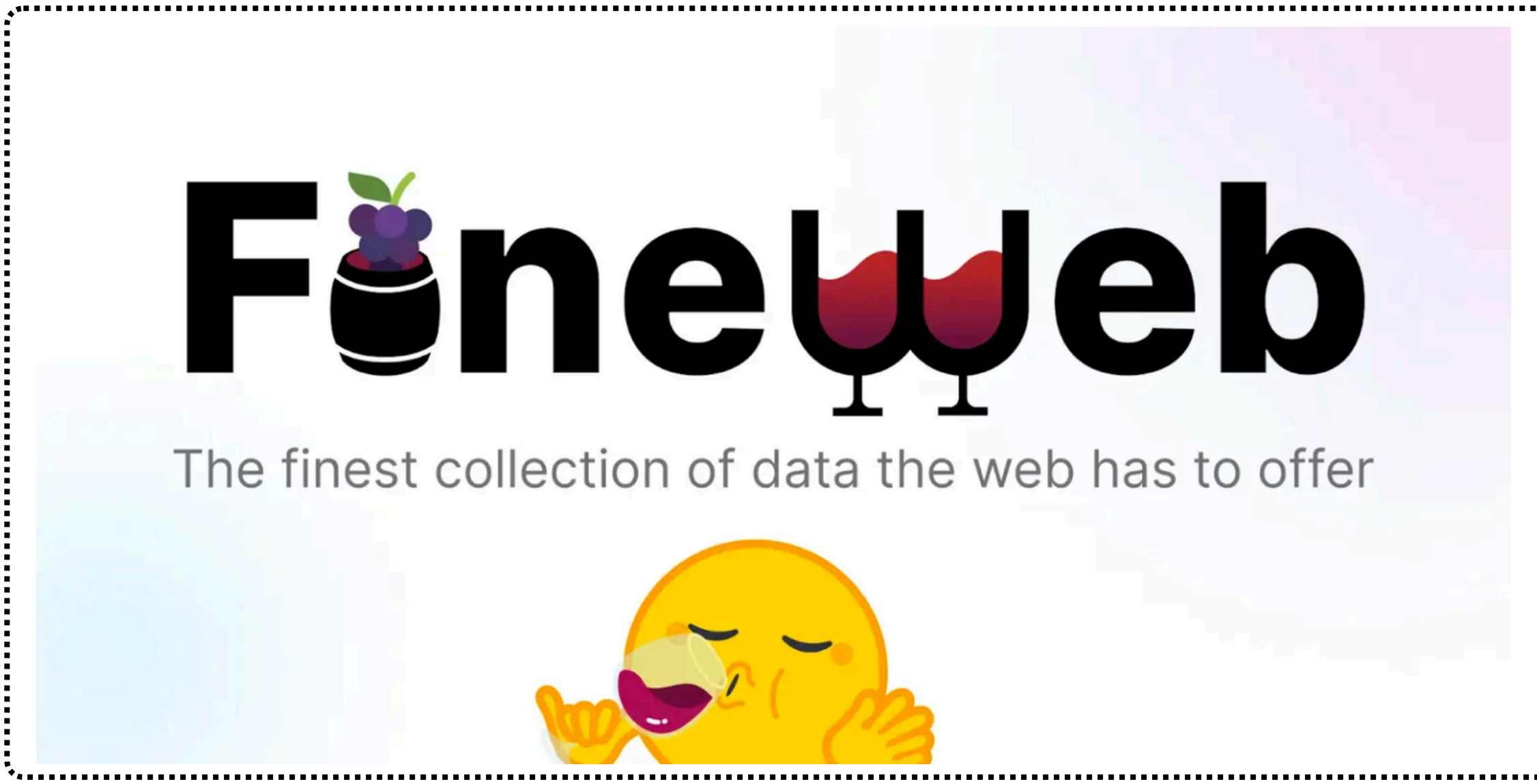
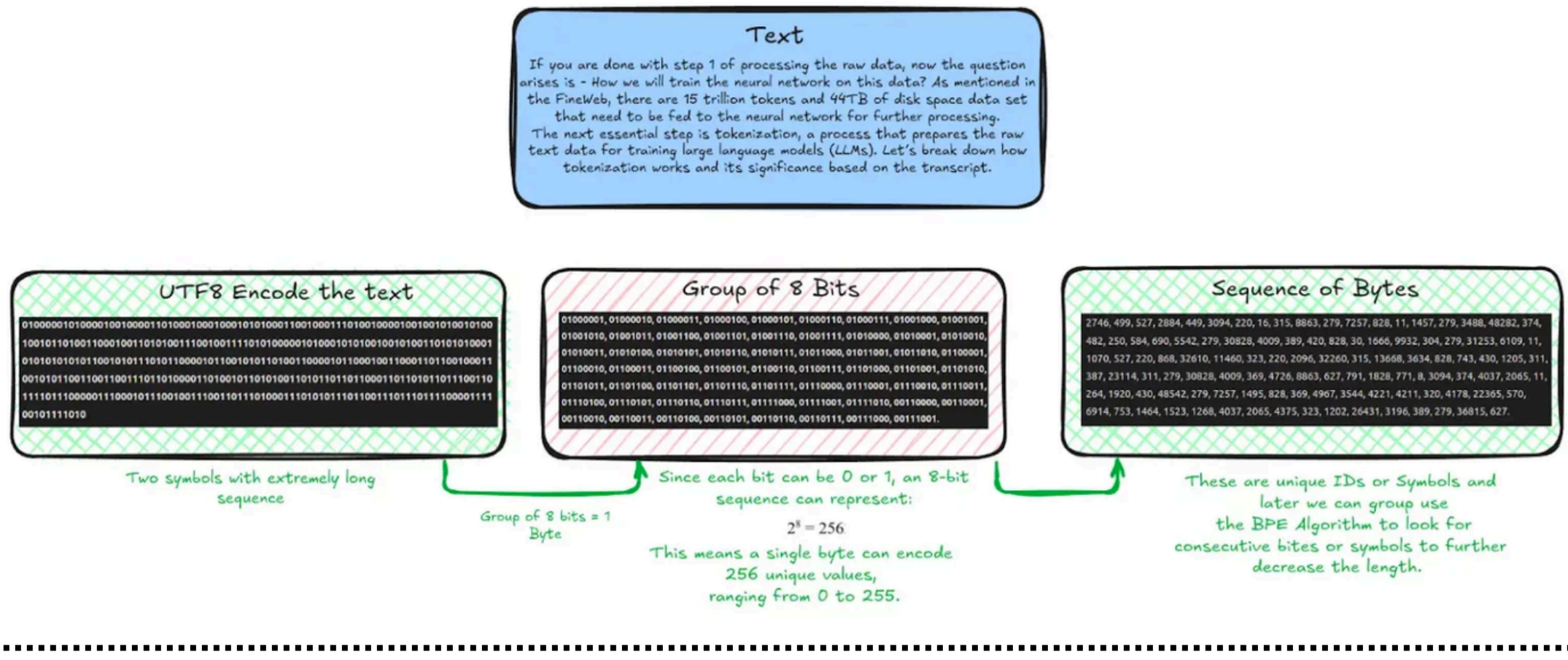
Analytics Vidhya

A Comprehensive Guide to Pre-training LLMs





What is the LLM Pretraining Stage?

- The LLM pretraining stage is the first phase of teaching a large language model (LLM) like how to understand and generate text.

- Think of it as reading a massive number of books, articles, and websites to learn grammar, facts, and common patterns in language.

- During this stage, the model processes billions of words (data) and predicts the next word (token) in a sentence repeatedly, refining its ability to generate coherent and relevant responses.

- However, at this point, it doesn’t fully “understand” meaning like a human—it just recognizes patterns and probabilities.

What can a Pre-trained LLM do?

- Pre-trained Large Language Models (LLMs) can perform a wide range of tasks, including text generation, summarization, translation, and sentiment analysis.

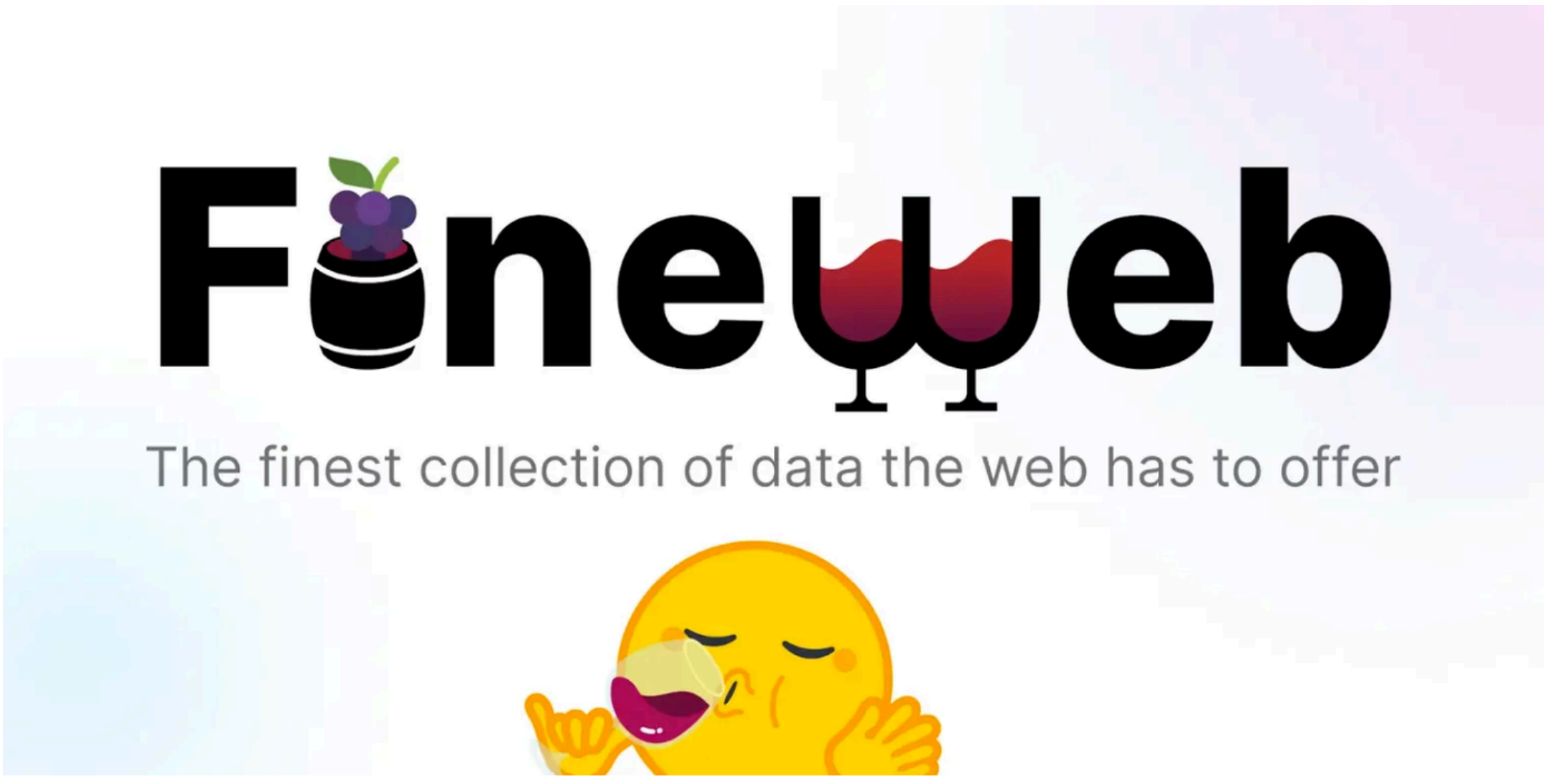
- They assist in code generation, question-answering, and content recommendation. LLMs can extract insights from unstructured data, facilitate chatbots, and automate customer support.

- They enhance creative writing, provide tutoring, and even generate realistic conversations. Additionally, they assist in data augmentation, legal analysis, and medical research by analyzing vast amounts of information efficiently.

- Their ability to understand and generate human-like text makes them valuable for various industries, from education and finance to healthcare and entertainment. However, they require fine-tuning for domain-specific accuracy.

Step 1: Process the Internet Data

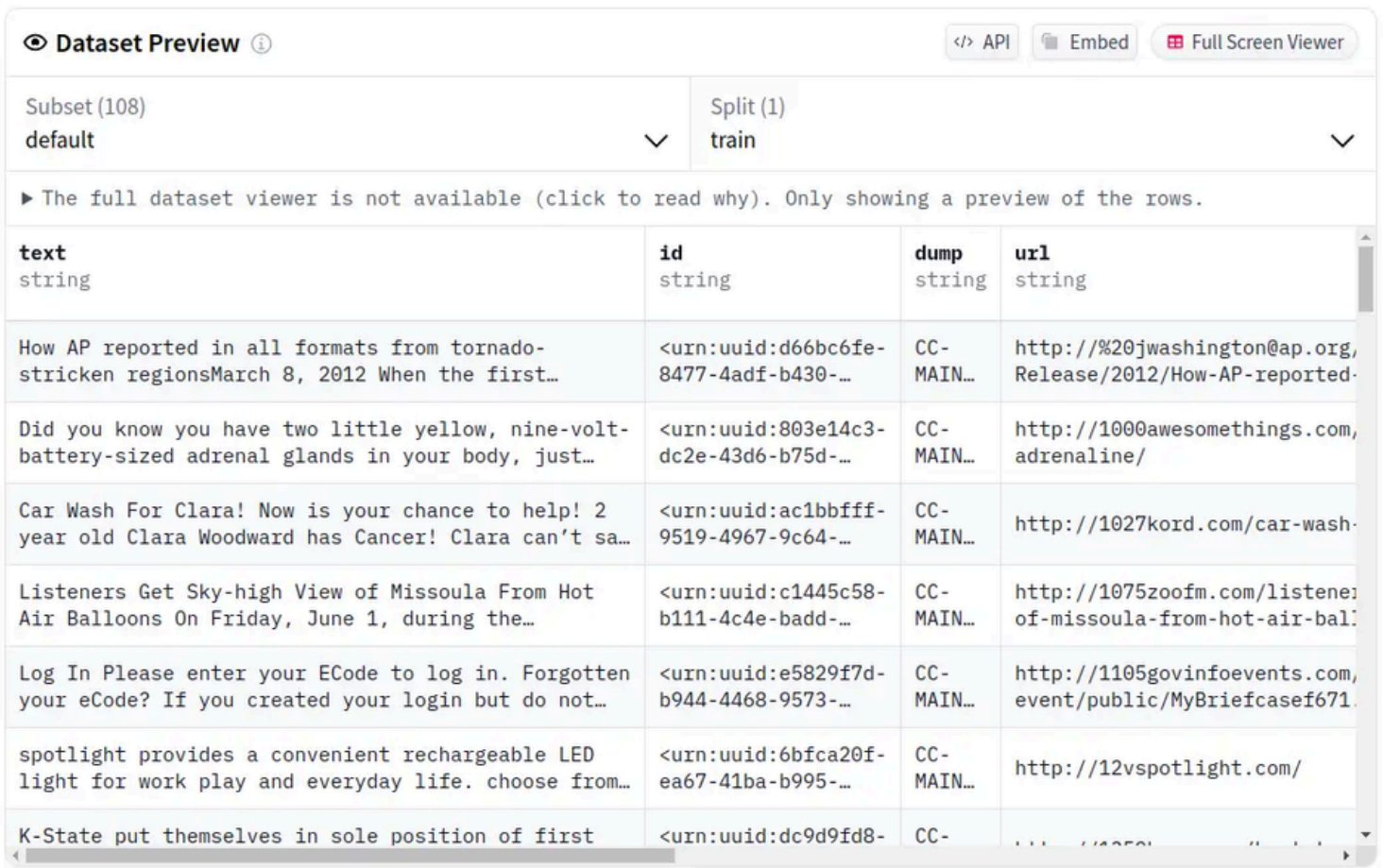
- There are multiple stages of training an LLM but here we will first talk about the LLM Pretraining stage.
- The performance of a large language model (LLM) is deeply influenced by the quality and scale of its pretraining dataset. If your dataset is clean, structured and easy to process, the model will work accordingly.



- However, for many state-of-the-art open LLMs like Llama 3 and Mixtral, the details of their pretraining data remain a mystery—these datasets are not publicly available, and little is known about how they were curated.

- To address this gap, Hugging collected data from the internet and curated FineWeb, a large-scale dataset (this is a portion of data available on the internet) specifically designed for LLM pretraining.

- This high-quality and diverse dataset has 15 trillion tokens and occupies 44TB of disk space, FineWeb is built from 96 CommonCrawl snapshots and has been shown to produce better-performing models than other publicly available pretraining datasets.



- What sets FineWeb apart is its transparency:

- It meticulously documented every design choice, running detailed ablations on deduplication and filtering strategies to refine the dataset's quality.

Where Does the Raw Data Come From?

There are two main sources:

- **Crawling the web yourself** – Used by companies like OpenAI and Anthropic.

- **Using public repositories** – CommonCrawl, a non-profit that has been archiving web data since 2007.

For FineWeb, they followed the approach of many LLM teams and used CommonCrawl (CC) as the starting point. CC releases a new dataset every 1-2 months, typically containing 200-400 TiB of text.

For example, the April 2024 crawl includes 2.7 billion web pages with 386 TiB of uncompressed HTML. Since 2013, CC has released 96 crawls, plus 3 older-format crawls from 2008-2012.

Analytics Vidhya

For more information, visit this article

Advanced | Best of Tech | Generative AI | Guide | LLMs

A Comprehensive Guide to Pre-training LLMs

Explore the fundamentals of LLM pretraining, its role in AI advancements, and how it shapes models like GPT-4.

Pankaj Singh 12 Feb, 2025