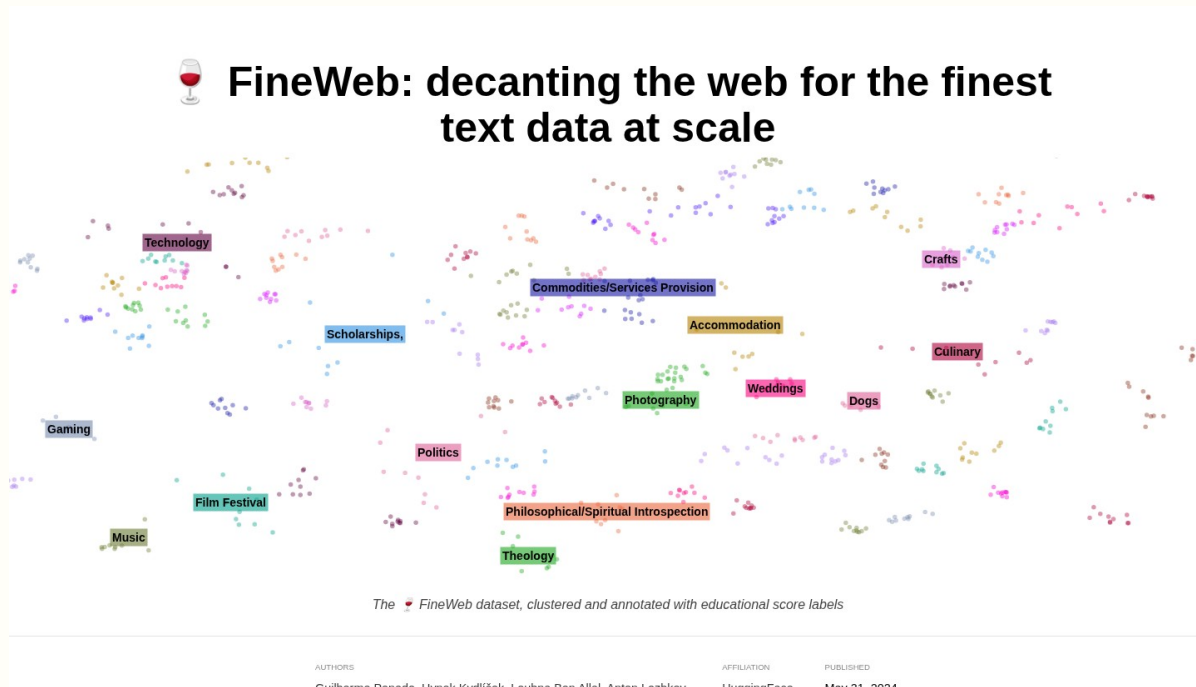


FineWeb

大規模言語モデルの事前学習用 Web データ

sasakiy84

公式サイト



1. FineWeb の目的と収録文書
2. Web 文書の収集・選別とデータ規模
3. データ品質の評価と話題分布
4. Web コーパスの利用動向と入手方法

1. FineWeb の目的と収録文書

事前学習データの公開を目的とする FineWeb

- FineWeb は、事前学習用の英語 Web データと処理コードを公開するために開発された [1]。
- 当時、Llama 3 などは重みを公開しても、学習データや収集・選別方法を十分に公開していなかった。
- そこで、先行する RefinedWeb の処理を再現・改良し、データ・比較用モデル・評価結果を公開した [1]。

言語モデルの開発工程と FineWeb の役割

段階	使うデータと目的
事前学習	大量の文章から言語・知識・表現を学ぶ
指示への調整	質問や依頼に沿う応答の例で調整する
能力の評価	別の問題を解かせ、学習後の能力を測る

FineWeb は事前学習に使う文章コーパスである [1]。

Web 文書を用いた次トークン予測

- 文章を細かなトークンへ区切り、前の部分から次のトークンの確率を学ぶ [1]。
- 実データの「Michigan unemployment claims workers ...」も、先頭から順に予測の対象になる。
- 配布 token_count は GPT-2 で数える [2]。

[1] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §2 <https://arxiv.org/abs/2406.17557>

[2] FineWeb / Data Fields <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

実例 1：雇用を報じるニュース

配布テキストからの抜粋（英語）

Michigan unemployment claims workers are losing their jobs.
NEW YORK (CNNMoney) -- Many jobless claims workers in the state of Michigan will soon be filing for unemployment themselves.

出来事や数量を述べる記事を学習する。同じ文書の後半には、関連記事や金利表も残っている [1]。

元ページ http://money.cnn.com/2012/08/03/news/economy/michigan-unemployment/index.htm?section=money_topstories

row 16 / <urn:uuid:67af6fe6-95e8-403c-821b-a7c16c3999bc> / 647 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

実例 2 : 調理の手順

配布テキストからの抜粋（英語）

You can also make one large shakshuka to share. Just spoon the sauce into a large ovenproof ceramic dish. Crack eggs into the dish, spaced evenly apart, then cook following the same recipe.

調理の手順を含む文書で、冒頭にはナビゲーションも残っている[1]。

元ページ <http://au.lifestyle.yahoo.com/better-homes-gardens/recipes/r/-/14032186/shakshuka-with-chorizo/>

row 27 / <urn:uuid:b967344e-3121-475b-9764-e76dd82eefbb> / 247 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

実例 3 : 登録案内も残る

配布テキストからの抜粋（英語）

You must be a registered member to view this page. |
If you are already a member, sign in now.
To register for your own account, sign up now.

元ページ <http://www.golivewire.com/forums/profile.cgi?p=539041910>

row 7 / <urn:uuid:528ab539-c6ba-441e-8389-65704e6c3be1> / 110 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

Web 記事の公開日とクロール日時

ニュースの実例に付く情報	値
収録クロール	CC-MAIN-2013-20
クロール日時	2013-05-21T10:29:46Z
文書全体の長さ	647 GPT-2 トークン

URL 中の記事日付は 2012-08-03 。収集日は公開日と異なり、抽出文には記事以外の表や関連記事も含まれる [1]。

[1] FineWeb / sample-10BT, row 16 <https://huggingface.co/datasets/HuggingFaceFW/fineweb/viewer/sample-10BT/train>

2. Web 文書の収集・選別とデータ規模

FineWeb のデータ作成工程

処理の段階	主な内容
Web 文書の収集	Common Crawl の HTML を使用
本文の抽出	本文抽出器 Trafilatura でテキスト化
言語・品質の選別	英語のスコアと反復・書式などで選別
文書の重複除去	クロール内で似た文書をまとめる
個人情報処理	メールと公開 IP アドレスを置換

元の HTML から本文を抽出する

- Common Crawl は、元の HTML を含む WARC と、抽出済みテキストの WET を公開している [1]。
- 原著では WET にメニューや定型文が多く残ることを確認し、WARC から Trafilatura という本文抽出器で取り出した。

英語の判定後に品質規則を適用する

- まず有害サイトなどの URL を除外し、言語識別器 fastText の英語スコアが 0.65 未満の文書を除く [1]。
- そこで、反復の多さや書式など、別の品質規則を重ねて適用する。

反復・書式に基づく文書の選別規則

- 同じ語句・行の繰り返し、短い行の多さ、句末記号の少なさなどを調べる [1]。
- 既存の Gopher ・ C4 の規則を調整し、さらに FineWeb 固有の規則を追加する。各規則の効果を学習実験で比べた。

MinHash による類似文書の重複除去

- 完全一致だけでは、見出しや一部の語が違う転載・再収録を取り除けない。文書の類似を近似判定する MinHash を使う [1]。
- 連続する 5 単語の並びを文書ごとに取り出し、その重なりを小さなハッシュの集合で比較する。
- 類似文書をまとめ、各グループから 1 件を残す。

クロール単位の重複除去を採用した理由

- 同じページが複数の年に収録されるため、全期間をまとめた重複除去も試した [1]。
- その結果、古いクロールに広告や不整形な文章が多く残った。頻繁に再収録される文章を消すと、分布も変わる [1]。
- そこで FineWeb は、クロールごとに独立して除去する方式を採用した。

メールアドレスと公開 IP の匿名化

- 公開時には、正規表現で見つけたメールアドレスと公開 IP アドレスを、用意した文字列へ置換した [1]。
- 電話番号は誤検出が多いため置換対象にしていない。

FineWeb の公開範囲とトークン数

範囲	クロール	トークン数
原著の FineWeb （2024）	96	約 15 兆
現行カードの FineWeb	110	約 18.5 兆
原著の FineWeb-Edu	—	約 1.3 兆

GPT-2 で算定。原著は 2013 ～ 2024-04、現行設定は 2025-06 まで。Edu は原著の 96 クロールから選別 [1, 2]。

[1] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. <https://arxiv.org/abs/2406.17557>

[2] 配布カード v1.4.0 / 2026-10-16 確認 <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

データ量・学習量・モデル規模の単位

表記	何を数えているか
15T トークン	配布データの規模。T は 1 兆
350B トークン	比較実験で学習に使う量。B は 10 億
1.71B パラメータ	モデルの学習可能な数値の個数

FineWeb-Edu の対象と選別基準

- FineWeb の文章を Llama 3 70B Instruct で採点し、その注釈を学習した軽量の回帰モデルで全体を選別する [1]。
- 0 ～ 5 の教育的価値スコアを使い、閾値 3 で約 1.3 兆トークンを残す。元の文章を選ぶ方式である。

[1] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4 <https://arxiv.org/abs/2406.17557>

[2] FineWeb2 公式リポジトリ <https://github.com/huggingface/fineweb-2>

教育的な価値を採点するモデルの学習

- 46 万文書を LLM で採点し、教育的な有用性の注釈を得る。原著は小中学校程度の知識を重視した [1]。
- 41 万文書で回帰モデルを学習し、残る 5 万文書で検証する。LLM の採点傾向を軽量なモデルで近似する。
- 全体への予測値を 0 ～ 5 の整数へ丸め、3 以上の文書を残す。これが FineWeb-Edu となる。

Web 由来の本文と LLM が生成した注釈

- FineWeb-Edu の本文は、元の Web 文書から選んだ文章である。
- LLM が生成するのは教育的な価値の注釈であり、その注釈に似た判定をする回帰モデルを学習する [1]。

3. データ品質の評価と話題分布

事前学習後の能力によるデータ品質の評価

- データを差し替えて同程度のモデルを事前学習し、外部の常識・知識・推論ベンチマークで比較する [1]。
- 処理の比較は主に約 1.8B パラメータ・約 28B トークン。大きな比較は 350B トークンを使う。

選択式問題の正解率と集約得点

説明用の設問	正解	選択結果	得点
水が凍る温度は？	0°C	0°C	1
地球の衛星は？	月	火星	0

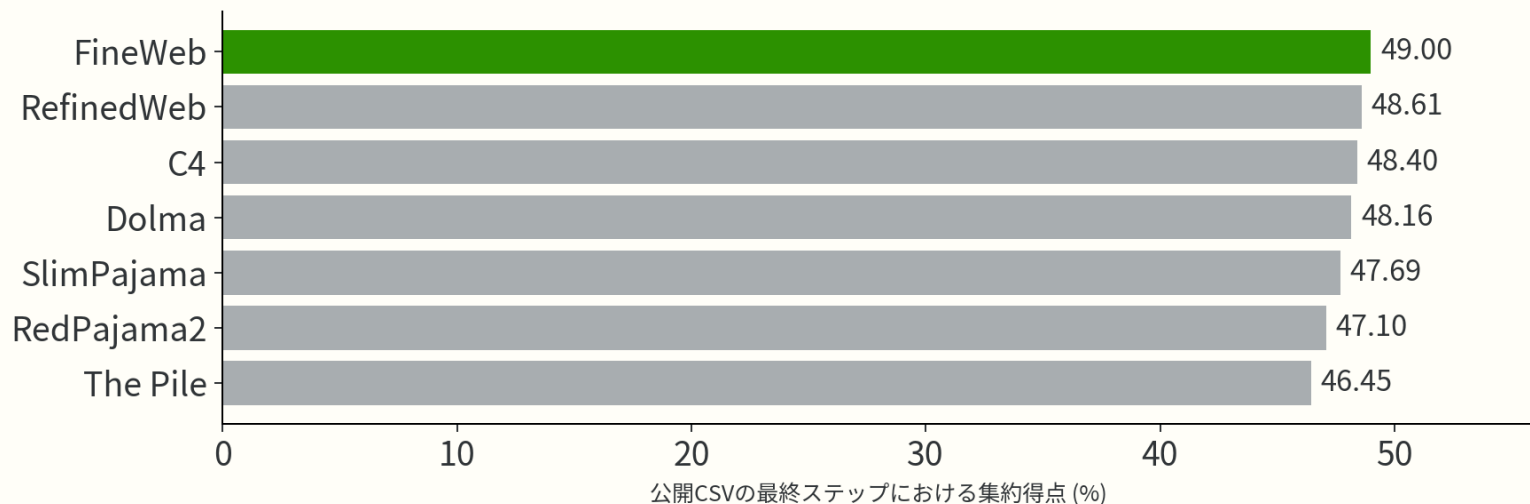
正解率 = 正解数 ÷ 設問数。この 2 問なら 50%。原著では選択肢全文の尤度や長さ正規化を用い、複数課題を集計する [1]。

原著で採用した知識・推論の評価課題

課題群	主に測るもの
HellaSwag / PIQA / SIQA	出来事の流れ・物理的 / 社会的常識
CommonSenseQA / WinoGrande	常識的な知識・文脈に応じた照応
ARC / OpenBookQA	科学の知識と推論
MMLU	多分野の知識問題

0-shot 評価。大きな課題は最大 1,000 例に制限する [1]。

公開 Web コーパスを用いたモデルの比較



公開 CSV の 167,000 step 時点。約 1.8B モデル、350B トークンを基準とする比較。小規模データは全量を使用 [1]。

[1] FineWeb eval_results.csv / 初期公開の比較系列 https://huggingface.co/datasets/HuggingFaceFW/fineweb/blob/main/eval_results.csv

FineWeb-Edu による MMLU ・ ARC の改善

原著 v2 の比較	FineWeb	FineWeb-Edu
MMLU の正解率	約 33%	約 37%
ARC の正解率	約 46%	約 57%

原著 v2 の 1.71B モデル ・ 350B トークンの比較 [1]。

原著の比較実験に用いたモデル構成

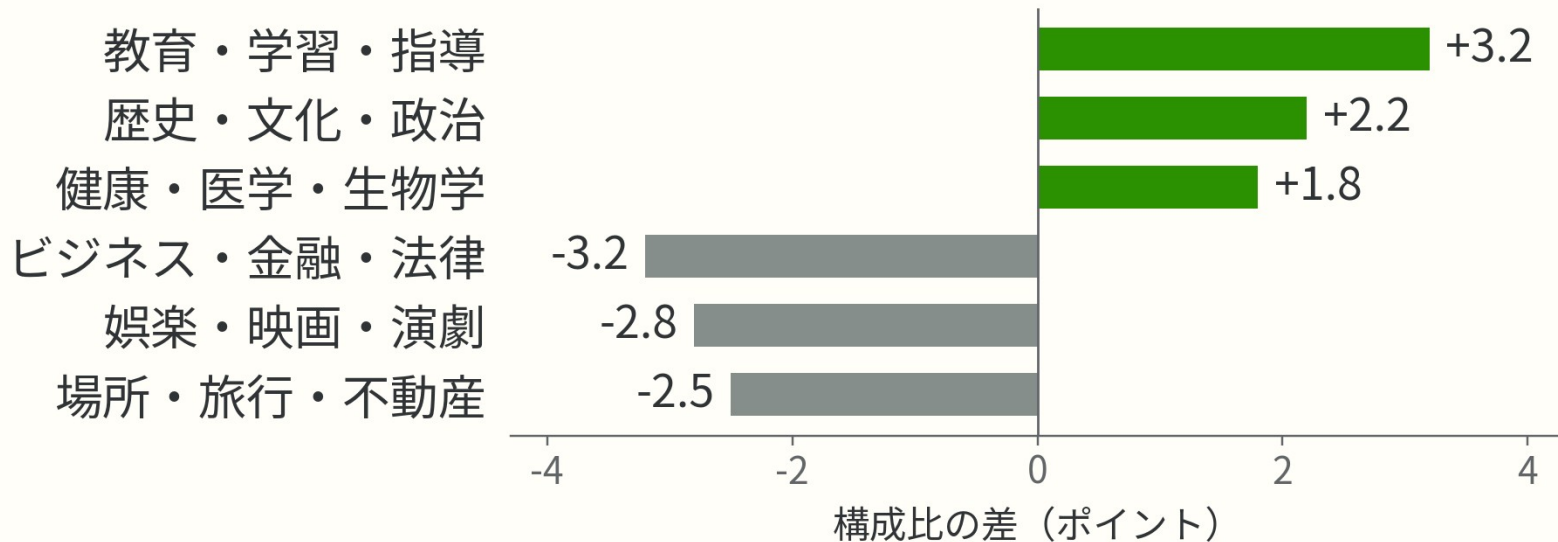
原著 v2 の構成	設定
モデル	Llama 型 / 1.71B パラメータ
文脈長 / 区切り方	2,048 トークン / GPT-2
比較する条件	実験内で学習量とステップ数を統一

データの選び方と初期値を変えた 2 モデルを学習し、得点を平均して偶然の影響を抑えた [1]。

選択肢の尤度に基づく回答の決定

- 評価では、問題文に対して各選択肢の全文がどれだけ続きやすいかを、モデルの尤度で比べる [1]。
- 短い選択肢が有利になる影響を抑えるため、課題に応じた長さの正規化も用いる。
- 正解率は、その方法で選んだ回答が正解と一致した割合である。

FineWeb-Edu で増減する話題の構成比



差は FineWeb-Edu – FineWeb。各 5 万文書の標本に基づく。
原著掲載 39 分類から、増加側・減少側の上位 3 分類ずつを例示 [1]。

話題の比較は各 5 万文書の自動分類に基づく

- FineWeb と FineWeb-Edu から各 5 万文書を抽出し、文の意味を数値化するモデルで文書を表現した [1]。
- 似た文書を自動でまとめ、密度の高い 100 クラスタを抽出し、Llama 3.1 70B で話題名を付けた。
- 図 18 は構成比の差の絶対値が 0.5 ポイント以上の 39 分類を示す [1]。

4. Web コーパスの利用動向と入手方法

公開 Web コーパスの登場年と開発目的

公開年	資源	FineWeb との関係
2020	C4	初期の大規模な英語 Web コーパス
2023	RefinedWeb	本文抽出・重複除去の先行研究
2024	FineWeb / Edu	処理の比較実験と全文データを公開
2025	FineWeb2	1,000 超の言語へ処理を適応

[1] Colin Raffel et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR. <https://arxiv.org/abs/1910.10683>

[2] Guilherme Penedo et al. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv. <https://arxiv.org/abs/2306.01116>

[3] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. <https://arxiv.org/abs/2406.17557>

[4] Hugging Face (2025), FineWeb 2 — README <https://github.com/huggingface/fineweb-2>

C4 ・ RefinedWeb ・ FineWeb2 の開発方針

- C4 （ 2020 ） は、言語モデル T5 の学習用に Common Crawl から構築した英語 Web コーパスである [1]。
- RefinedWeb （ 2023 ） は、選別・重複除去した Web 文書で学習し、書籍等を混ぜたデータとの性能差を検証した [2]。
- FineWeb2 （ 2025 ） は、言語ごとに選別条件を調整し、多言語コーパスへ拡張した [3]。

[1] Colin Raffel et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR. <https://arxiv.org/abs/1910.10683>

[2] Guilherme Penedo et al. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv. <https://arxiv.org/abs/2306.01116>

[3] FineWeb2 — README <https://github.com/huggingface/fineweb-2>

FineWeb 系列を利用した公開モデル

- SmolLM （2024）は、重複除去した FineWeb-Edu の 220B トークンを、合成教材・Python コードと混合した [1]。
- SmolLM3 （2025）の公式記事には、FineWeb-Edu での構成比較と、FineWeb2 を含む事前学習の記録がある [2]。

[1] SmolLM 公式記事 <https://huggingface.co/blog/smollm>

[2] SmolLM3 公式記事 <https://huggingface.co/blog/smollm3>

SmolLM（2024）の事前学習コーパス

使用したデータ	トークン数
重複除去した FineWeb-Edu	220B
合成教材・物語の Cosmopedia v2	28B
教育的な Python コード	4B

B は 10 億トークンを表す [1]。

SmolLM3 （ 2025 ） の事前学習の混合比

学習の段階	Web	コード	数学
第 1 段階	85%	12%	3%
第 2 段階	75%	15%	10%
第 3 段階	63%	24%	13%

Web 部分に FineWeb-Edu ・ DCLM ・ FineWeb2 等を使用。各値はその段階の学習データ全体に対する割合 [1]。

クロール別・トークン量別の配布設定

配布設定	対象
default	公開されている全クロール
CC-MAIN-2024-10 など	指定したクロール
sample-10BT など	小規模に取り出したトークン量のサンプル

10BT ・ 100BT ・ 350BT は含まれる文書が重なる [1]。

[1] FineWeb / Smaller sample versions <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

FineWeb のストリーミング読み込み方法

```
from datasets import load_dataset

fw = load_dataset(
    "HuggingFaceFW/fineweb", name="sample-10BT",
    split="train", streaming=True,
)
row = next(iter(fw))
print(row["text"])
```

10BT は約 100 億 GPT-2 トークンの設定。ストリーミングで順次読み込める [1]。

データセットの利用条件と原文の権利

- データセットは ODC-By 1.0 で公開され、Common Crawl の利用規約にも従う [1]。
- データベースのライセンスと、各 Web 文書の著作権・利用条件は分けて確認する。

FineWeb の目的・規模と教育的価値による選別

- 英語 Web の本文を選別・重複除去し、次トークン予測のための大規模な事前学習データとして公開した。
- データと処理の比較実験を公開した点が特徴。初版 15 兆、現行カードは約 18.5 兆トークンを示す。
- FineWeb-Edu は教育的価値で文書を選別し、知識・推論の得点と話題の構成比を変える。

[1] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. <https://arxiv.org/abs/2406.17557>

[2] FineWeb data card <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

Appendix

1. 配布文書の構造と長さの集計
2. 収録された Web 文書の追加例
3. 文書選別・話題分布と評価実験の詳細
4. 配布データの仕様と開発体制
5. 読後の研究に関連する文献

文書と出所のフィールド

フィールド	内容
text	抽出・整形された Web 文書
id / dump	Common Crawl の文書 ID / クロール名
url / date	元ページの URL / クロール日時
file_path	元 WARC ファイルへのパス

自動生成されたメタデータ

フィールド	意味
language	配布対象は英語（en）
language_score	fastText による言語予測スコア
token_count	GPT-2 tokenizer によるトークン数

FineWeb-Edu には教育的価値の score / int_score も付く [1, 2]。

[1] FineWeb data card <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

[2] FineWeb-Edu data card <https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>

100 文書の部分集合の分析

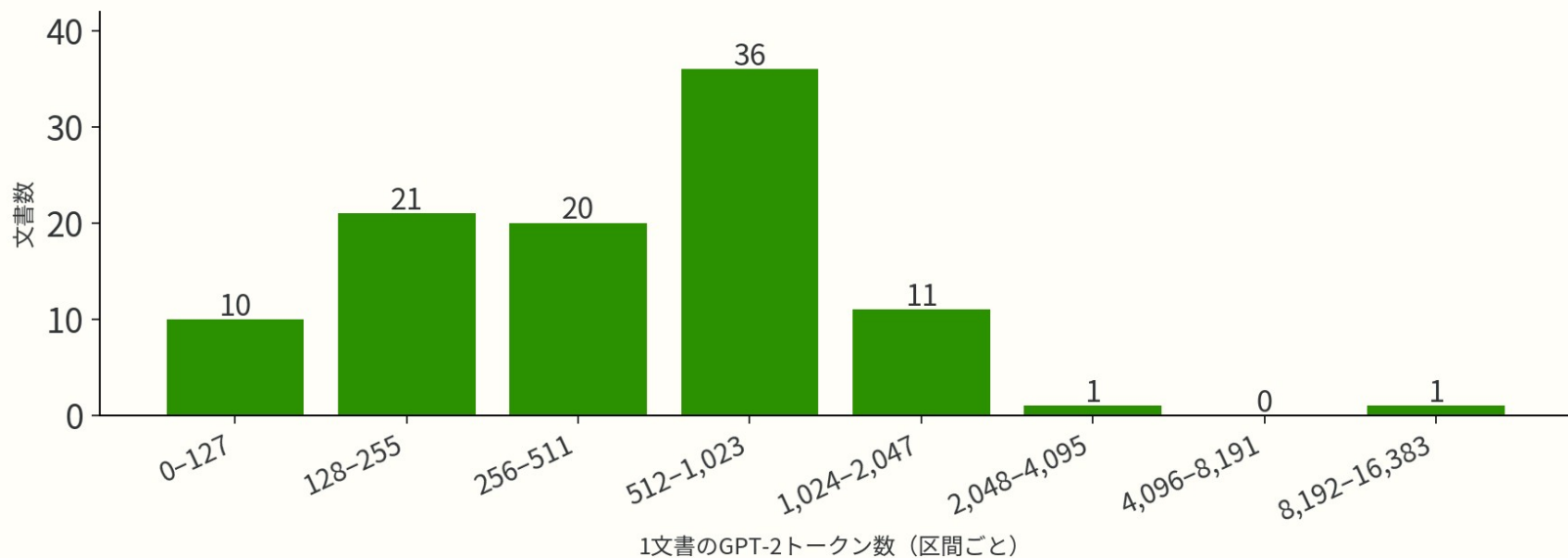
sample-10BT の先頭 100 文書を対象とする
収録クローलは CC-MAIN-2013-20
文書の長さと、収録された文章の例を確認する

先頭 100 文書のトークン数の記述統計

GPT-2 トークン数	値
平均 / 中央値	649.02 / 471.5
母標準偏差	974.63
第 1 / 第 3 四分位点	210.5 / 736.0
最小 / 最大	73 / 8,990

先頭 100 件の配布値を集計。分位点は線形補間。

先頭 100 文書のトークン数の分布



最短の例：装飾品へのレビュー

配布テキストからの抜粋（英語）

Great decorating addition

I have a grape/Italian theme in my kitchen. I purchased 5 of these. I decided to use them to put around my pull knobs on my overhead cabinets.

この 100 件では最短の 73 トークン [1]。

元ページ <http://www.yankeecandle.com/detail/wine-collection/1262116>

row 13 / <urn:uuid:424e4cce-c33a-49b4-b148-37c8a973ba5c> / 73 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

最長の例：人物紹介と多数の投稿

配布テキストからの抜粋（英語）

Larry Kim, Founder and Chief Technology Officer
Larry Kim founded WordStream in 2007.

この 100 件で最長の 8,990 トークン。人物紹介の後に、別々の投稿の見出し・概要が多数続く [1]。

元ページ <http://www.wordstream.com/larry-kim>

row 38 / <urn:uuid:07f0d7eb-c5f6-4de5-b87d-ffedbc8204de> / 8,990 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 1 : テレビドラマの掲示板

配布テキストからの抜粋（英語）

| Viewing Single Post From: Spoilers for the Week of February 11th |
| Lil | | Feb 1 2013, 09:58 AM |
Don't care about Chloe/Taniel/Jen-Jen. Don't care about Sami, really, but hoping that we get some good "SAMANTHA GENE!!" Marlena Death-Stares out of it.

感情的な発言、固有名、顔文字が含まれる。会話的な Web 文も事前学習の対象になる [1]。

元ページ <http://daytimeroyaltyonline.com/single/?p=8906650&t=8780053>

row 0 / <urn:uuid:39147604-bf8e-4ed5-b19c-54105f8ae8a7> / 142 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 2：医療ニュースの一覧

配布テキストからの抜粋（英語）

A novel two-step immunotherapy approach has shown clinically beneficial responses in patients with advanced ovarian cancer.
Following Lifestyle Recommendations Reduces Risk of Cancer Death

複数の見出し・記事概要が連続している [1]。

元ページ <http://news.cancerconnect.com/>

row 2 / <urn:uuid:07b8e00d-b445-4736-a593-cd1c147dce21> / 576 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 3：共有とごみ削減のブログ

配布テキストからの抜粋（英語）

Free the Cans! Working Together to Reduce Waste
In a blog about how people share, it's worth the occasional reference to the bizarre ways that people DON'T SHARE.

社会的な話題を扱う意見文。筆者の評価や問いかけを含む文章も混ざる [1]。

元ページ <http://sharingsolution.com/2009/05/23/free-the-cans-working-together-to-reduce-waste/>

row 3 / <urn:uuid:c970d9a2-a5ce-4050-9ea3-58d7bbd609a8> / 575 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 4：大学のスポーツニュース

配布テキストからの抜粋（英語）

September 28, 2010

2010 Season - Bowman pulls down CCIW honor

| Matt Bowman was named CCIW "Runner of the Week" after his fourth place finish at the Brissman-Lundeen Invitational in Rock Island, Illinois on September 24. |

大学の選手表彰について、日付・人名・大会名を伴って報じる文章 [1]。

元ページ <http://www.augustana.edu/x22236.xml>

row 5 / <urn:uuid:a42b6d10-fba9-4678-af39-5b242fdfe790> / 626 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 5：食品業界の新商品記事

配布テキストからの抜粋（英語）

Kraft Foods has taken the Cadbury chocolate brand in a new direction, by combining it with cheese for the first time.

チョコレートとチーズを組み合わせた商品について、ブランド名や発売の説明が載る [1]。

元ページ <http://www.fdin.org.uk/2012/01/kraft-launches-new-philadelphia-with-cadbury/>

row 6 / <urn:uuid:8ba5fa5a-1f92-4f5a-95e3-85dbb7befbfe> / 337 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 6：飲食店の衛生検査記録

配布テキストからの抜粋（英語）

```
| Facility Type: | | Full Service Restaurant |  
| Inspection date: | | March 27, 2012 |  
| Number of critical violations: | | 3 |  
| Number of non-critical violations: | | 3 |
```

表を抽出した縦棒や、違反件数・検査日が残る。自然な段落文以外の構造もある [1]。

元ページ http://www.healthspace.com/Clients/VDH/Fairfax/Fairfax_Website.nsf/Food-InspectionDetails?OpenView&R...

row 8 / <urn:uuid:4ad1d60f-e986-44b4-8b58-e833a6dddba8> / 786 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 7 : アトリエのイベント案内

配布テキストからの抜粋（英語）

News of the Week
Barrie Spring Studio Tour
April 27th & 28th
10:00 til 4:00 pm

日時・場所・販売の案内を含む。局地的・時限的な情報も Web データの一部になる [1]。

元ページ <http://www.jillpricestudios.ca/artist/website/size/big>

row 9 / <urn:uuid:7bc4fe2c-c860-45d0-91f4-7cb5d3800e66> / 415 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 8：医療費不正への注意喚起

配布テキストからの抜粋（英語）

Tips for Preventing Medicare Fraud

The Department of Health and Human Services (HHS) Office of the Inspector General (OIG) has created a new web site to provide tips for preventing Medicare fraud and medical identity theft. See HHS news release.

行政の公開情報を紹介する短文。リンク先への案内が主な内容になる文書もある [1]。

元ページ <http://aota.org/Consumers/Ethics/Fraud.aspx>

row 19 / <urn:uuid:957006ac-2ad8-43d0-8919-52d5a4373d10> / 91 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 9 : Perl のプロファイラの不具合調査

配布テキストからの抜粋（英語）

I've a new release of NYTProf ready to upload but I'm stuck.
The CPAN Testers service is reporting a failure on a number of systems
but I can't reproduce it locally or work out the cause.
Can you reproduce the failure with Devel::NYTProf 2.07_94?

プログラムのバージョン、再現条件、調査への協力依頼を含む技術的な文章 [1]。

元ページ <http://blog.timbunce.org/2009/01/28/can-you-reproduce-this-nytprof-failure/?like=1&wpnonce=9501e873bf>

row 21 / <urn:uuid:095340b9-8a67-45ac-afb3-250bec8e3efd> / 351 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

追加例 10 : 楽譜の販売ページ

配布テキストからの抜粋（英語）

Reprinted from the renowned Bach-Gesellschaft edition, this work features the complete Sonatas and Partitas for Unaccompanied Violin and the six Sonatas for Violin and Clavier.

楽譜の紹介と購入に関する文章 [1]。

元ページ <http://store.doverpublications.com/0486236838.html>

row 22 / <urn:uuid:3cad8aa8-a0c4-4556-81a3-03253c1cb027> / 568 GPT-2 tokens

[1] FineWeb / sample-10BT <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

WARC と WET の形式と本文抽出方法

- WARC は HTML やリクエストのメタデータを保持する。WET は抽出済みテキストである [1]。
- 原著は WET にメニューなどが多く残ることを確認し、WARC から Trafilatura で本文を抽出した。

MinHash の設定と近似の意味

- 文書を英単語の 5-gram へ分け、112 個のハッシュを 14 組 × 8 個にまとめる [1]。
- どれかの組で 8 個が一致した文書を結び、推移的なクラスタを作る。各クラスタから 1 件を残す。
- 約 75% 以上似た文書を狙う確率的な方式である [1]。

教育スコアの学習用注釈

原著 v2 の設定	値・方法
LLM による採点	460,000 文書、0 ～ 5 点
学習 / 検証	410,000 / 50,000 文書
回帰モデル	Snowflake-arctic-embed-m + 線形層
閾値 3 での F1	82% (LLM 注釈を正解とする)

原著は 460k 注釈、初期カードは 500k 表記 [1, 2]。

[1] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4 <https://arxiv.org/abs/2406.17557>

[2] FineWeb-Edu data card <https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>

初期公開 CSV と原著 v2 の評価条件

- 公開 CSV は 7 データセットの学習曲線を含む初期の比較系列。最終 167,000 step の agg_score を図に使用した [1]。
- FineWeb-Edu の MMLU ・ ARC 比較には、原著 v2 本文の丸め値を使用した [2]。

[1] eval_results.csv https://huggingface.co/datasets/HuggingFaceFW/fineweb/blob/main/eval_results.csv

[2] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4 <https://arxiv.org/abs/2406.17557>

FineWeb-Edu 側で多い話題（ 1/3 ）

話題（原著ラベルの和訳）	差（ポイント）
教育・学習・指導	+3.2
歴史・文化・政治	+2.2
健康・医学・生物学	+1.8
エネルギー・環境・気候	+1.2
水・環境・保全	+1.2
歴史・王族・哲学	+1.2

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 6 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb-Edu 側で多い話題（2/3）

話題（原著ラベルの和訳）	差（ポイント）
野生生物・動物・自然	+1.2
宇宙・天文学・科学	+1.1
歴史・宗教・文化	+1.0
魚・環境・保全	+0.9
歴史・戦争・軍事	+0.9
天気・気候・科学	+0.8

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 6 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb-Edu 側で多い話題（3/3）

話題（原著ラベルの和訳）	差（ポイント）
植物・園芸・自然	+0.8
健康・栄養・食生活	+0.7
文化・インド・精神性	+0.6
健康・医学・疾患	+0.5
数学・数式・教育	+0.5
数学・教育・指導	+0.5

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 6 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb 側で多い話題（1/4）

話題（原著ラベルの和訳）	差（ポイント）
ビジネス・金融・法律	-3.2
娯楽・映画・演劇	-2.8
場所・旅行・不動産	-2.5
食・夏・レシピ	-2.1
スポーツ・チーム・試合	-2.1
音楽・娯楽・芸術	-2.1

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 6 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb 側で多い話題（2/4）

話題（原著ラベルの和訳）	差（ポイント）
個人生活・家族・余暇	-1.4
ファッション・衣服・アクセサリー	-1.3
犯罪・法律・警察	-1.2
ゲーム・技術・ゲーム作品	-1.2
技術・ガジェット・革新	-1.1
スポーツ・フットボール・サッカー	-1.0

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 6 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb 側で多い話題（3/4）

話題（原著ラベルの和訳）	差（ポイント）
自動車・自動車産業・産業	-1.0
文学・戦争・ファンタジー	-0.8
手芸・個人生活・日記	-0.8
家具・建築・デザイン	-0.8
宝飾品・アクセサリー・買い物	-0.8
政治・紛争・国際情勢	-0.7

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 6 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb 側で多い話題（4/4）

話題（原著ラベルの和訳）	差（ポイント）
野球・スポーツ・ニュース	-0.6
技術・サポート・マーケティング	-0.6
デート・人間関係・場所	-0.5

各 5 万文書。差 = FineWeb-Edu - FineWeb。絶対差 0.5 ポイント以上の 39 分類から 3 分類 [1]。
[1] Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. §4.1, Fig. 18. <https://arxiv.org/abs/2406.17557v2>

FineWeb-Edu の話題分布と適合する分野

- 原著の話題分析では、FineWeb と Edu から各 50,000 件を埋め込み、100 個の密なクラスタを比較した [1]。
- Edu は教育・歴史などの比重が高く、商業・娯楽・旅行などの比重が低い。
- 次トークンの予測しにくさを表す perplexity でも得意分野が異なり、一般 Web では FineWeb 側が有利な分野もある。

FineWeb の公開版と主な更新内容

版 / 公開日	主な変更
v1.0.0 / 2024-04-21	初期公開
v1.1.0 / 2024-05-31	11 クロールの重複除去を修正・4 月分を追加
v1.3.0 / 2025-01-31	未処理文書を修正、特定ドメインを削除
v1.4.0 / 2025-07-11	2025 年 1 ～ 6 月の 6 クロールを追加

現行カードの変更履歴から抜粋 [1]。

[1] FineWeb / Changelog <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

配布されているファイルサイズの概要

対象	公表・取得メタデータの容量
現行 FineWeb 全体	カード掲載値 約 50.45 TB
sample-10BT	15 Parquet ファイル / 約 30.64 GB
実際に分析した対象	viewer から取得した 100 文書

sample-10BT : 30,639,384,917 bytes (10 進 GB)。カード掲載値は 27.6 GB [1, 2]。

[1] FineWeb data card <https://huggingface.co/datasets/HuggingFaceFW/fineweb>

[2] sample/10BT ファイル一覧 / 2026-10-16 <https://huggingface.co/datasets/HuggingFaceFW/fineweb/tree/main/sample/10BT>

配布設定とリビジョンを指定した読込

```
from datasets import load_dataset

fw = load_dataset(
    "HuggingFaceFW/fineweb",
    name="sample-10BT", split="train", streaming=True,
    revision="9bb295ddab0e05d785b879661af7260fed5140fc",
)
row = next(iter(fw))
```

FineWeb 系列の対象言語と選別基準

資源	対象と位置付け
FineWeb	英語 Web の幅広い事前学習用文書
FineWeb-Edu	教育的価値で選んだ英語文書
FineWeb-Edu-score-2	閾値 2 で残す量を増やした派生版
FineWeb2	多言語向けに言語判定・規則を調整

FineWeb2 は言語ごとに全体を重複除去する [2]。

[1] FineWeb-Edu 公式カード <https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>

[2] FineWeb2 公式リポジトリ <https://github.com/huggingface/fineweb-2>

FineWeb の開発者と関連する活動

- Guilherme Penedo が第一著者、Thomas Wolf が最終著者。原著の所属表記は Hugging Face [1]。
- 両者は大規模テキスト処理ライブラリ datatrove の公開引用にも共著者として記載されている [2]。
- Hynek Kydlíček、Loubna Ben Allal、Anton Lozhkov、Margaret Mitchell、Colin Raffel、Leandro von Werra との共同研究。

[1] Guilherme Penedo et al. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS. <https://arxiv.org/abs/2406.17557>

[2] datatrove README / Citation <https://github.com/huggingface/datatrove>

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel et al. / 2020 / Journal of Machine Learning Research 21(140), 1-67

T5 の原著。英語 Web コーパス C4 と、テキスト変換に統一した学習・評価を提案する。

<https://arxiv.org/abs/1910.10683>

The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only

Guilherme Penedo et al. / 2023 / arXiv:2306.01116

Falcon の事前学習用コーパスを報告し、Web データの選別と重複除去の効果を検証する。

<https://arxiv.org/abs/2306.01116>

The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale

Guilherme Penedo et al. / 2024 / NeurIPS, Datasets and Benchmarks

FineWeb と Edu の原著。抽出・重複除去・品質選別の比較実験を報告する。

<https://arxiv.org/abs/2406.17557>

FineWeb: decanting the web for the finest text data at scale

Hugging Face / 2024-05-31 / 公式技術記事

処理の選択と失敗した手法を、学習曲線や具体例とともに記録する。

<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

FineWeb

HuggingFaceFW / Hugging Face Datasets / 2026-10-16 参照

公式データカード。配布設定・更新履歴・容量・フィールド・利用条件を記載する。

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

FineWeb-Edu: the Finest Collection of Educational Content

Anton Lozhkov et al. / 2024 / Hugging Face

教育的価値による選別、分類器の学習、閾値別の派生版を説明する。

<https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>

SmolLM - blazingly fast and remarkably powerful

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch / 2024 / Hugging Face Blog

FineWeb-Edu を合成教材とコードへ混ぜた、小規模言語モデルの開発記録。

<https://huggingface.co/blog/smollm>

SmolLM3: smol, multilingual, long-context reasoner

Hugging Face / 2025 / Hugging Face Blog

公開データの混合、構成比較、多言語・長文脈対応を含むモデル開発を説明する。

<https://huggingface.co/blog/smollm3>

FineWeb2: One Pipeline to Scale Them All -- Adapting Pre-Training Data Processing to Every Language

Guilherme Penedo et al. / 2025 / arXiv:2506.20920

FineWeb の処理を多言語に広げ、言語別の選別・重複除去を検証する。

<https://arxiv.org/abs/2506.20920>