

Datasets: monology/**pile-uncopyrighted**

like 121

Modalities: Text

Formats: json

Size: 1M - 10M

ArXiv: arxiv:2101.00027

Libraries: Datasets

Dask

Croissant + 1

License: other

Dataset card

Viewer

Files

Community 8

Dataset Viewer (First 5GB)

Auto-converted to Parquet

API

Embed

Full Screen Viewer

Split (3)

train · 892k rows

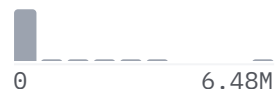


Search this dataset

SQL Console

text

string · *lengths*



meta

dict

It is done, and submitted. You can play "Survival of the Tastiest" on Android, and on the web...

```
{ "pile_set_name": "Pile-CC" }
```

```
<?xml version="1.0" encoding="UTF-8"?> <segment>
<name>PD1</name> <description>Patient Additional...
```

```
{ "pile_set_name": "Github" }
```

Topic: reinvent midnight madness Amazon announced a new service at the AWS re:Invent Midnight Madness...

```
{ "pile_set_name": "Pile-CC" }
```

About Grand Slam Fishing Charters As a family owned business we know how important it is that your tri...

```
{ "pile_set_name": "Pile-CC" }
```

Downloads last month **150,784**

Use this dataset

Edit dataset card



Size of the auto-converted Parquet files (First 5GB per split):
3.65 GB

Number of rows (First 5GB per split):
1,252,122

Models trained or fine-tuned on monolog...

MiniLLM/MiniPLM-llama3.1-212M

Text Generation • Updat... • 908 • 1

MiniLLM/Ref-Pretrain-Qwen-104M

End of preview. [Expand](#) in  Dataset Viewer.

< Previous **1** 2 3 ... 8,918 Next >

Pile Uncopyrighted

In response to [authors demanding that LLMs stop using their works](#), here's a copy of [The Pile](#) with all copyrighted content removed.

Please consider using this dataset to train your future LLMs, to respect authors and abide by copyright law.


Creating an uncopyrighted version of a larger dataset (ie RedPajama) is planned, with no ETA.

Methodology


Cleaning was performed by removing everything from the Books3, BookCorpus2, OpenSubtitles, YTSubtitles, and OWT2 subsets.

Based on section 7.1 of [the original paper](#), these datasets are the only ones which are not explicitly allowed to be used in AI training.


 Text Generation • Updat... • ⬇ 878 • ❤ 1

 MiniLLM/MiniPLM-Qwen-200M


 Text Generation • Updated Oct ... • ⬇ 265

 MiniLLM/MiniPLM-Qwen-500M

 Text Generation • Updat... • ⬇ 182 • ❤ 5

 MiniLLM/Pretrain-Qwen-200M

 Text Generation • Updated Oct ... • ⬇ 176

 MiniLLM/Pretrain-Qwen-500M

 Text Generation • Updated Oct ... • ⬇ 147

[Browse 20 models trained on this dataset](#)

 System theme

TOS

Privacy

About

Jobs



Models

Datasets

Spaces

Pricing

Docs