SmolLM3-3B

Public Summary of Training Content

Summary v1.0: https://huggingface.co/spaces/hfmlsoc/smollm3-eu-data-transparency/commits/main - Last updated: 25/07/2025

This Space contains the transparency report for the SmolLM3-3B: https://huggingface.co/HuggingFaceTB/SmolLM3-3B GPAI model developed by Hugging Face: https://huggingface.co/ following the guidelines provided by the AI Office.

It may serve as an example for **open-source GPAI trained exclusively on public datasets**. For more information, see the Explanatory Notice and Template: https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models

📋 TL;DR

SmolLM3-3B is a state-of-the-art 3-billion parameter language model by **Hugging Face** trained on **10+ trillion tokens** from publicly available datasets including web documents, scientific articles, and code. Training focused on **6 EU languages** plus others. The model uses **only public datasets** (no commercial licensing, user data, or other private data). Data processing was done by the original component dataset curators with **varied approaches to TDM and filtering** that typically include compliance with robots.txt and other opt-out mechanisms, and educational content classifiers.

1. General information

TL;DR: Provider: Hugging Face | Model: SmolLM3-3B | Training: 10+ trillion tokens, 6 EU languages + others

👇 Click for full information ▼

1.1. Provider identification

- **Provider name and contact details:**
  - Hugging Face Inc., Email: legal@huggingface.co

1.2. Model identification

- **Versioned model name(s):**

- **SmolLM3-3B**

- **Model dependencies:**
  - **None**

1.3. Modalities, overall training data size and other characteristics

- **TEXT**
  - **Size: [more than 10 trillion tokens]**
    - The training corpus is made up of 11 trillion tokens as tokenized by the Llama-3.2-1B: https://huggingface.co/meta-llama/Llama-3.2-1B tokenizer.
  - The training corpus for SmolLM3 is made up of several publicly accessible large datasets containing web documents, scientific articles, software code, and synthetically generated textbooks and mathematical data for pre-training in addition to several mid-training and fine-tuning datasets to enable chat interactions, instruction-following and task-solving behaviors.
- **Latest date of data acquisition/collection for model training:**
  - The training dataset is made up of different subsets with different publication and cutoff dates. For pre-training, the earliest dataset was last updated on 4/3/2024 (Stack v2), and the latest on 2/19/2025 (FineWeb2-HQ)
- **Description of the linguistic characteristics of the overall training data:**
  - The overall training process focuses on 6 languages that are all official EU languages: English, French, Spanish, German, Italian, and Portuguese. In addition, pre-training intentionally included smaller quantities of data in Mandarin Chinese, Russian, Persian, Japanese, Korean, Vietnamese, Hindi, Thai, and Greek (also an official EU language). Other languages may have been included due to the limitations of automatic language identification in filtering stages.
- **Other relevant characteristics of the overall training data:**

- The training data also includes software code in the programming languages included in the Stack v2 dataset (16 languages including C, C++, C-Sharp, Python, Java, JavaScript, Markdown, HTML, Shell, PHP, TypeScript, Swift, SQL, Ruby, Go, and Rust).

2. List of data sources

TL;DR: ✅ Publicly available datasets including synthetic data | ❌ No commercial licensing, crawling, user data, or private data

👇 Click for full information ▼

2.1. Publicly available datasets

- **Have you used publicly available datasets to train the model?**
  - ☑ **Yes**
- **If yes, specify the modality(ies) of the content covered by the datasets concerned:**
  - ☑ **Text**
- **List of large publicly available datasets:**
  - DCLM: https://hf.co/datasets/mlfoundations/dclm-baseline-1.0: https://hf.co/datasets/mlfoundations/dclm-baseline-1.0
  - FineWeb-Edu: https://hf.co/datasets/HuggingFaceFW/fineweb-edu: https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu
  - FineWeb2: https://hf.co/datasets/epfml/FineWeb2-HQ: https://huggingface.co/datasets/epfml/FineWeb2-HQ
  - Stack V2: https://hf.co/datasets/bigcode/the-stack-v2:

https://hf.co/datasets/bigcode/the-stack-v2

- pes2o: https://hf.co/datasets/allenai/peS2o: https://hf.co/datasets/allenai/peS2o
- FineMath : https://hf.co/datasets/HuggingFaceTB/finemath: https://hf.co/datasets/HuggingFaceTB/finemath
- MegaMath: https://hf.co/datasets/LLM360/MegaMath: https://hf.co/datasets/LLM360/MegaMath
- SmolTalk2: https://hf.co/datasets/HuggingFaceTB/smoltalk2: https://huggingface.co/datasets/HuggingFaceTB/smoltalk2

- **General description of other publicly available datasets not listed above:**
  - In addition to the large datasets cited above, many additional publicly available datasets were added to target specific domains, including several math datasets made up of both web-filtered and synthetic data, Wikipedia data, "reasoning data" generated by selected large models on diverse problems, Jupyter notebooks for code, and synthetically generated textbooks; all in English language or software code. The full list of pre-training datasets is available at the following URL: https://hf.co/collections/HuggingFaceTB/smollm3-pretraining-datasets-685a7353fdc01aecde51b1d9: https://hf.co/collections/HuggingFaceTB/smollm3-pretraining-datasets-685a7353fdc01aecde51b1d9

> ## 2.2. Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

- **Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?**
  - ☐ **No**

2.2.2. Private datasets obtained from other third parties

.

- Have you obtained private datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?
  - ☐ No

2.3. Data crawled and scraped from online sources

- Were crawlers used by the provider or on behalf of?
  - ☐ No

2.4. User data

- Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?
  - ☐ No
- Was data collected from user interactions with the provider's other services or products used to train the model?
  - ☐ No

2.5. Synthetic data

- Was synthetic AI-generated data created by the provider or on their behalf to train the model?
  - ☑ Yes
- If yes, modality of the synthetic data:
  - ☑ Text
- Information about other AI models, including provider's own AI model(s) not available on the market, used to generate synthetic

data to train the model to which this Summary applies:

- Additional data was generated using the Qwen3-32B: https://huggingface.co/Qwen/Qwen3-32B and Qwen3-0.6B: https://huggingface.co/Qwen/Qwen3-Embedding-0.6B open-weight models.

- **Additional comments (optional):**

  - Some parts of the SmolTalk2: https://hf.co/datasets/HuggingFaceTB/smoltalk2 dataset mentioned in Section 2.1 were synthetically generated for the purpose of training this model. See License information in the dataset description: https://huggingface.co/datasets/HuggingFaceTB/smoltalk2#license.

2.6. Other sources of data

- **Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?**

  - ☐ **No**

3. Data processing aspects

TL;DR: TDM rights: robots.txt baseline otherwise dataset-dependent | Content filtering: Dataset-dependent including educational classifiers

👇 Click for full information ▼

3.1. Respect of reservation of rights from text and data mining exception or limitation

- **Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:**
    - The training corpus for SmolLM3-3B is made up of diverse pre-existing public datasets maintained by various organizations who still have their own approach to managing the TDM exception. All crawl-based data in the datasets uses the CommonCrawl archives which comply with robots.txt. Some datasets such as the Stack v2 additionally offer general opt-out mechanisms. For each dataset, the latest publicly available version was used to ensure propagation of any rights reservation expressed to the dataset custodian.

3.2. Removal of illegal content

- **General description of measures taken:**
    - Each of the component datasets leveraged is the product of a distinct curation effort by its custodians to select the most desirable content. The specific approaches can typically be found in the dataset documentation. Among other factors, most of the datasets take the approach of using classifiers to identify "highly educational" samples that lowers the likelihood of illegal content. Datasets like The Stack v2 for example additionally filtered out software code with copyleft licenses and applied automatic redaction of categories of personal data.