



Open and responsible development of
Large Language Models for code

Leandro von Werra | Machine Learning Engineer @ Hugging Face | @lvwerra
Harm de Vries | Lead of the LLM lab @ ServiceNow | @harmdevries77

GITHUB COPILOT: CHAT



GitHub Copilot

Hi @monalisa, how can I help you?

I'm powered by AI, so surprises and mistakes are possible. Make sure to verify any generated code or suggestions, and share feedback so that we can learn and improve.

Ask a question or type '/' for commands



parse_expenses.py × addresses.rb × sentiments.ts ×

```
1 import datetime
2
3 |
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
```

Closed model APIs



Open model weights



Closed model APIs

 Model weights not available

- Can't run the model locally
- Can't inspect the model's representations
- Limits fine-tuning abilities

Open model weights

Closed model APIs

Model weights not available

- Can't run the model locally
- Can't inspect the model's representations
- Limits fine-tuning abilities

Open model weights

Training data is not disclosed

- Content creators don't know if their data is used and there's no way to remove it
- Can't inspect data for biases
- Potential benchmark contamination
- Limits scientific reproducibility

BigCode: open-scientific collaboration

We are building LLMs for code in a collaborative way:

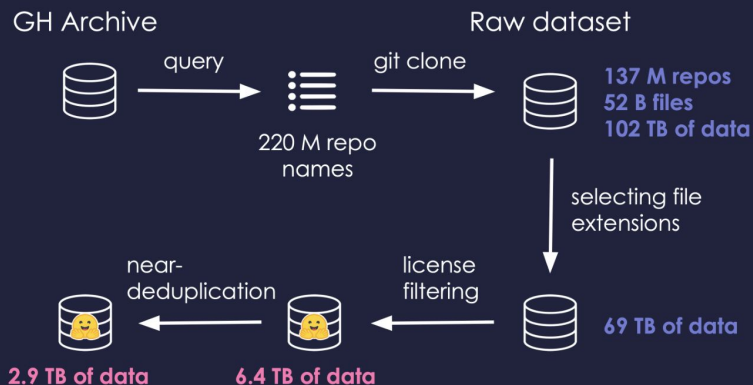
- Full data transparency
- Open source processing and training code
- Model weights released with commercial friendly license

1100+ researchers,
engineers, lawyers, and
policy makers



The Stack

Data collection



Find the filtered and deduplicated datasets at: www.hf.co/bigcode

Data inspection + Opt-out



BigCode

The Stack is an open governance interface between the AI community and the open source community.

Am I in The Stack?

As part of the BigCode project, we released and maintain [The Stack](#), a 3.1 TB dataset of permissively licensed source code in 30 programming languages. One of our goals in this project is to give people agency over their source code by letting them decide whether or not it should be used to develop and evaluate machine learning models, as we acknowledge that not all developers may wish to have their data used for that purpose.

This tool lets you check if a repository under a given username is part of The Stack dataset. Would you like to have your data removed from future versions of The Stack? You can opt-out following the instructions [here](#).

The Stack version:

v1.1
▼

Your GitHub username:

StarCoder

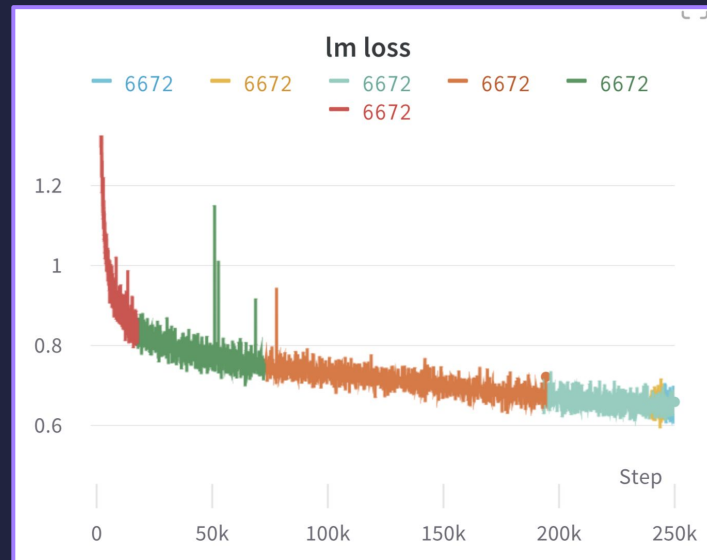
Model size: 15B parameters

Context length: 8096 tokens

Infrastructure: 512 GPUs

Training length: 1T tokens / 250k steps

Training time: 24 days

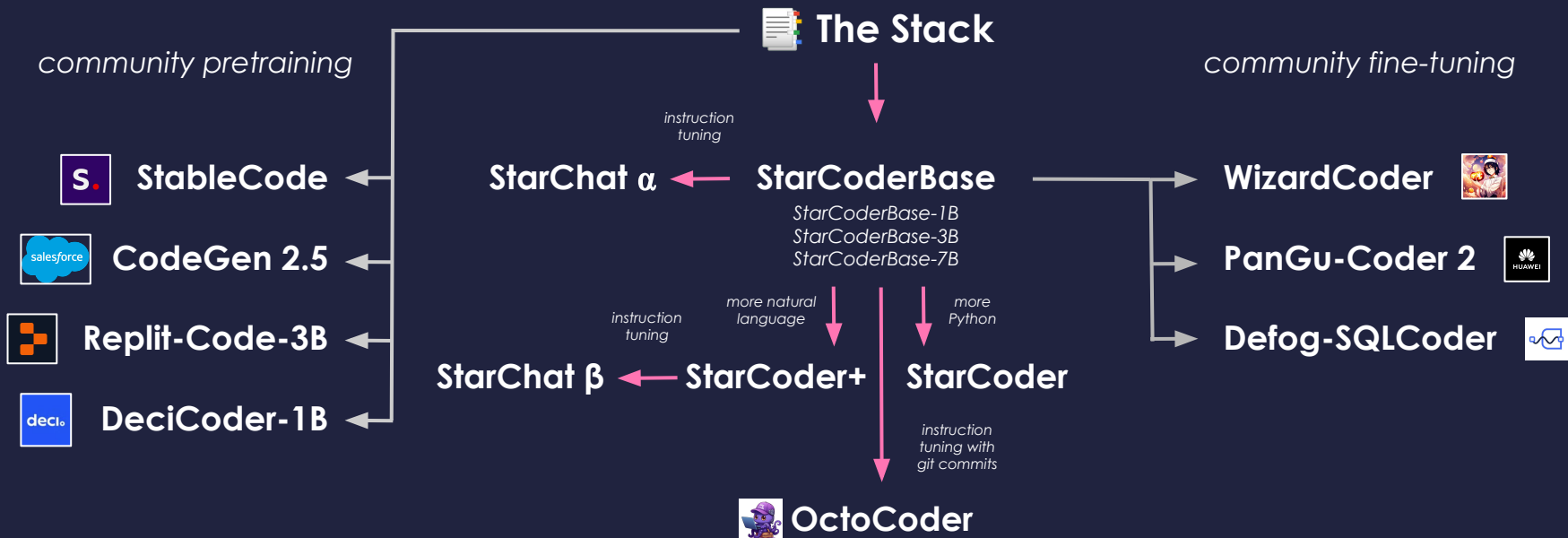


“smooth sailing”

Best open LLM for code at the time of release!



BigCode Ecosystem





StarCoder2



nVIDIA

×



×

servicenow[®]

Partnership with



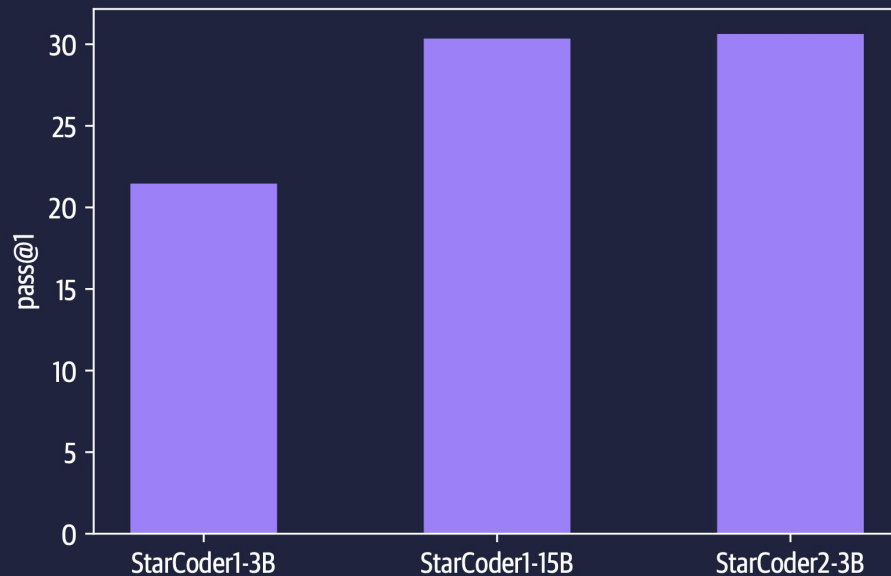
Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

1. Alignment of values and vision
 - a. <https://www.softwareheritage.org/2023/10/19/swh-statement-on-llm-for-code/>
2. Ensure long term availability of the training set via the archive
3. Share processing scripts such as the deduplication pipeline
4. Ease traceability via SWHID

StarCoder2: First promising results!

- **StarCoder2-3B** is on par with **StarCoder-15B**
- **5x smaller** → **5x faster/cheaper** inference



Thank you! And come join us!



Questions?

www.bigcode-project.org

hf.co/bigcode