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
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.
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

Sep 2022
The Stack

Data collection

GH Archive

query

220 M repo names

2.9 TB of data

near-deduplication

6.4 TB of data

Raw dataset

git clone

137 M repos
52 B files
102 TB of data

selecting file extensions

69 TB of data

Data inspection + Opt-out

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

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:

Check!
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

Community pretraining:
- StableCode
- CodeGen 2.5
- Replit-Code-3B
- DeciCoder-1B

The Stack:
- StarChat α
- StarCoderBase
  - StarCoderBase-1B
  - StarCoderBase-3B
  - StarCoderBase-7B
- StarCoder+
- StarCoder
- OctoCoder

Instruction tuning:
- StarChat α
- StarCoderBase
- StarCoder

More natural language:
- StarCoder

More Python:
- StarCoder

Community fine-tuning:
- WizardCoder
- PanGu-Coder 2
- Defog-QLCoder
StarCoder2
Partnership with

1. Alignment of values and vision
   a. https://www.softwareheritage.org/2023/10/19/swh-statement-on-lm-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