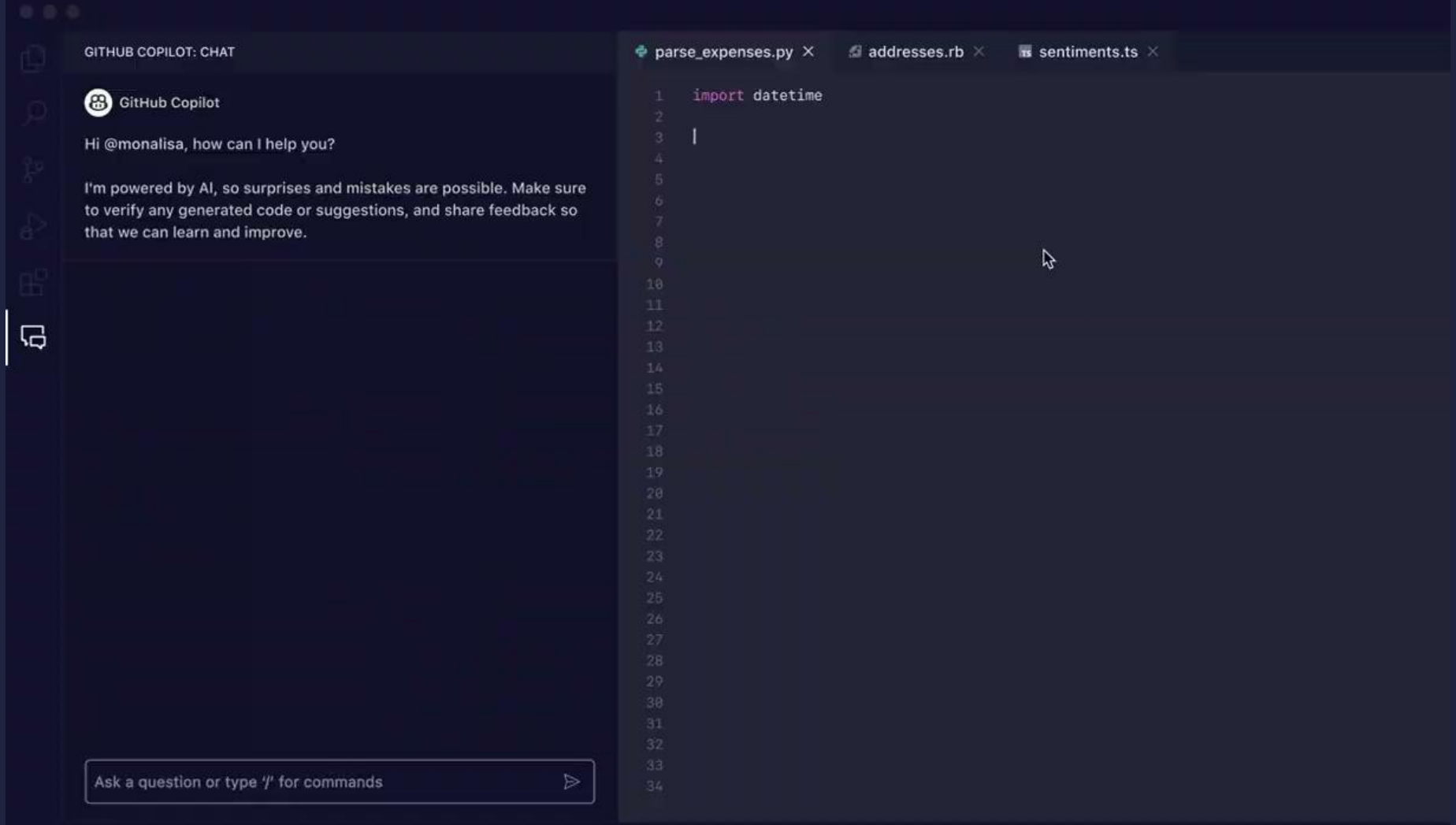
BigCode

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
import datetime

Closed model APIs

OpenAI

Gemini

ANTHROP\C

Open model weights

MistralAI

deepseek

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

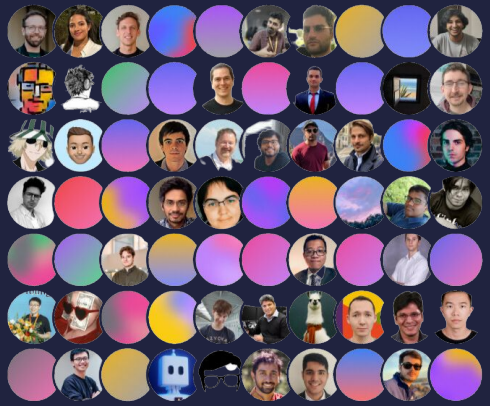
Sep 2022

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

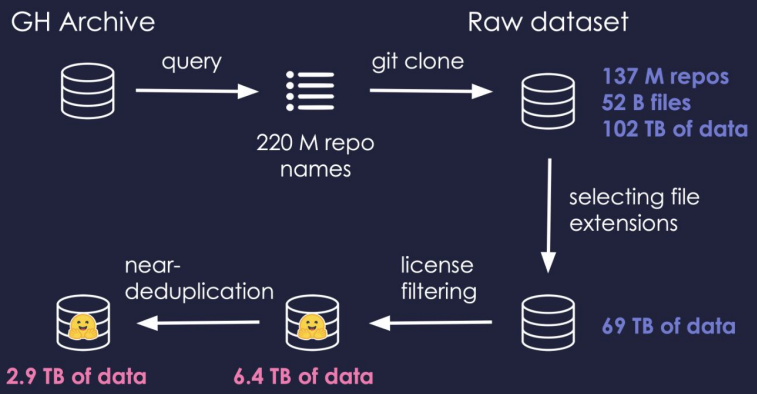
1100+ researchers, engineers, lawyers, and policy makers



Nov 2022

The Stack

Data collection



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

Data inspection + Opt-out

May 2023

StarCoder

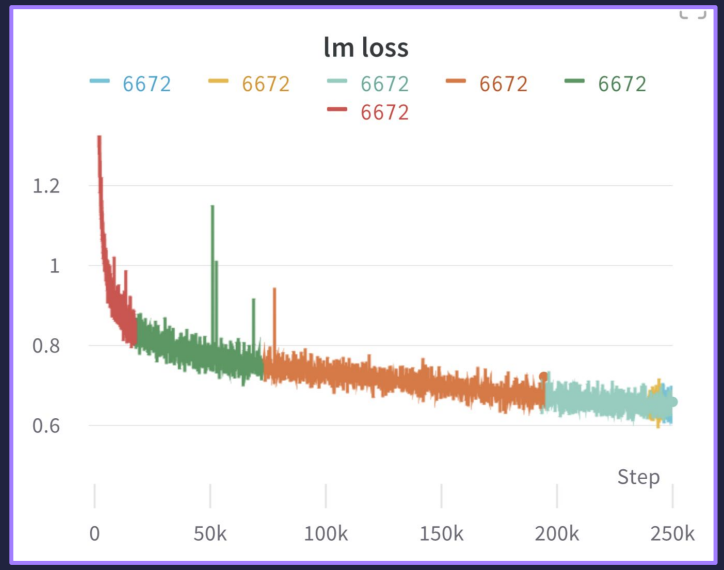
Model size: 15B parameters

Context length: 8096 tokens

Infrastructure: 512 GPUs

Training length: 1T tokens / 250k steps

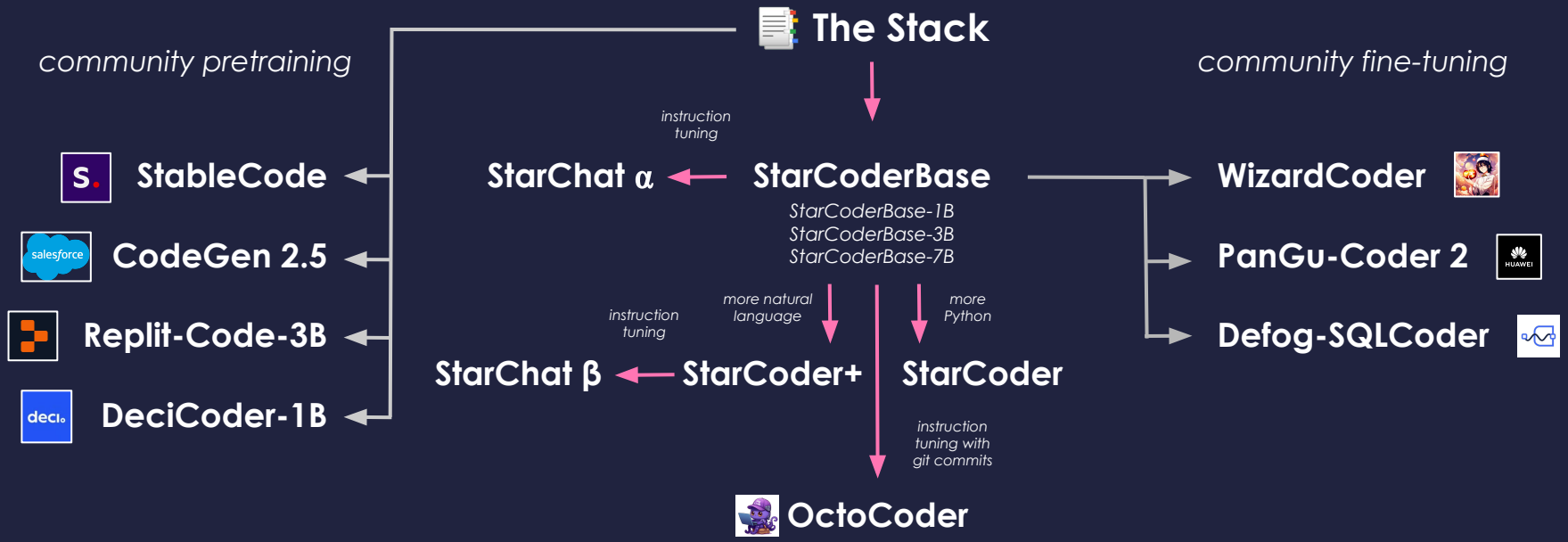
Training time: 24 days



"smooth sailing"

Best open LLM for code at the time of release!

BigCode Ecosystem




StarCoder2
NVIDIA
×
×
servicenow

Partnership with

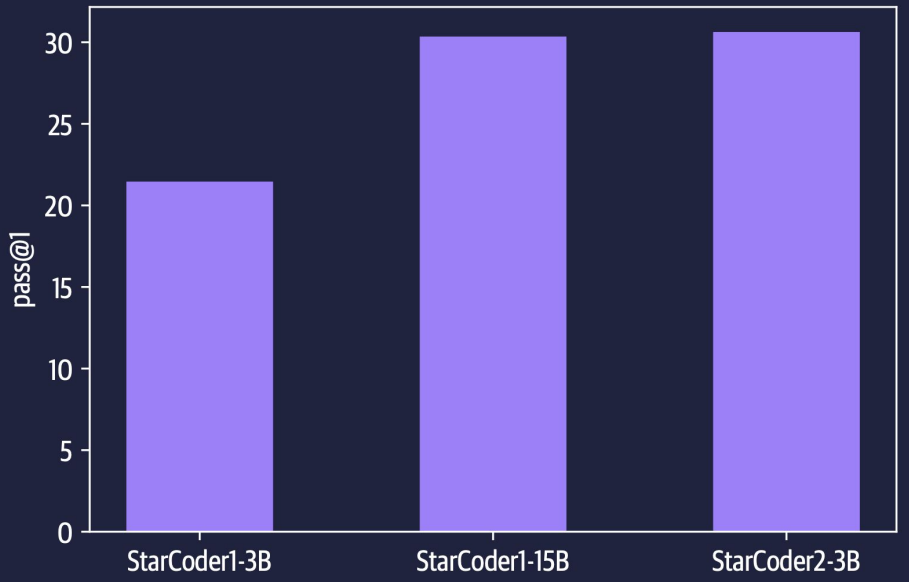
Software Heritage

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