

Open Large Language Models for Code

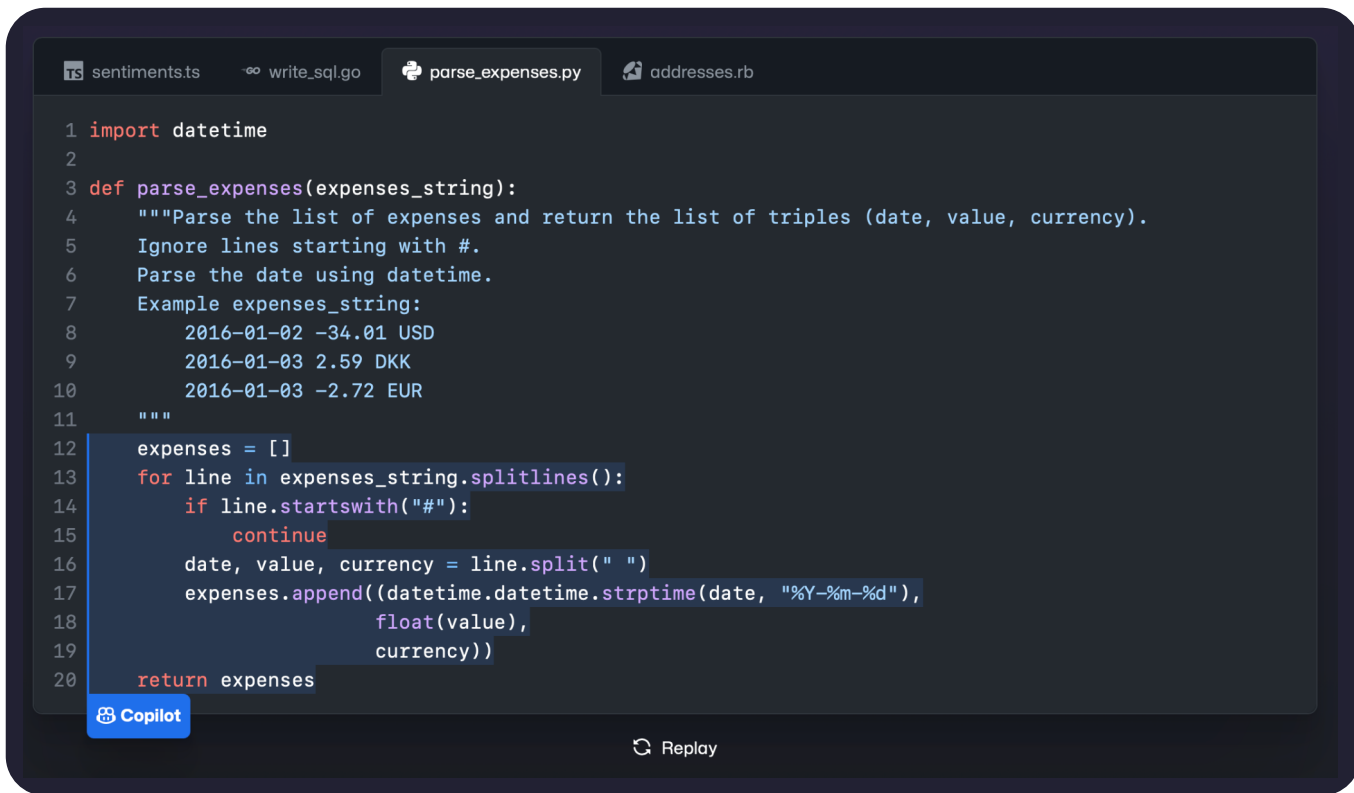
Loubna Ben Allal

Machine Learning Engineer, Science team




Hugging Face


How it started: GitHub Copilot in 2021



```
ts sentiments.ts  ∞ write_sql.go  parse_expenses.py  addresses.rb

1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

 Copilot

 Replay

How it's going: Over 1.7k open models trained on code



Search models, datasets, users...

Models Datasets Spaces Posts Docs Solutions Pricing



Tasks Libraries Datasets Languages Licenses Other

Models 1,738

Filter by name

new Full-text search

Sort: Trending

code

Reset Languages

code x

stabilityai/stable-code-instruct-3b

Text Generation • Updated 2 days ago • 2.21k • 95

stabilityai/stable-code-3b

Text Generation • Updated 11 days ago • 20.1k • 590

microsoft/phi-2

Text Generation • Updated Feb 6 • 740k • 3k

bigscience/bloom

Text Generation • Updated Jul 28, 2023 • 11k • 4.5k

PipableAI/pip-library-etl-1.3b

Text Generation • Updated 1 day ago • 830 • 19

Locutusque/OpenCerebrum-1.0-7b-DPO

Text Generation • Updated 2 days ago • 158 • 11

microsoft/phi-1_5

Text Generation • Updated Feb 6 • 81.2k • 1.25k

bigcode/starcoder

Text Generation • Updated 9 days ago • 14.9k • 2.68k

m-a-p/OpenCodeInterpreter-DS-33B

Text Generation • Updated 27 days ago • 1.91k • 64

bigcode/starcoder2-15b

Text Generation • Updated 24 days ago • 128k • 471

bartowski/stable-code-instruct-3b-GGUF

Text Generation • Updated 5 days ago • 7

aurora-m/aurora-m-biden-harris-redteamed

Text Generation • Updated 2 days ago • 395 • 13

VAIBHAV22334455/JARVIS

Text Generation • Updated 1 day ago • 164 • 6

codellama/CodeLlama-7b-hf

Text Generation • Updated Jan 29 • 152k • 275

How did we get here?



Strong Instruction-tuned and **base models**

★ Big Code Models Leaderboard

Inspired from the 🤖 [Open LLM Leaderboard](#) and 🤖 [Open LLM-Perf Leaderboard](#), we compare performance of base multilingual code generation models on [HumanEval](#) benchmark and [MultiPL-E](#). We also measure throughput and provide information about the models. We only compare open pre-trained multilingual code models, that people can start from as base models for their trainings.

🔍 Evaluation table

📊 Performance Plot

📄 About

🚀 Submit results

📄 See All Columns

🔍 Search for your model and press ENTER...

⌵ Filter model types

all

base

instruction-tuned

T	▲	Model	▲	Win Rate	▲	humaneval-python	▲	java	▲	javascript	▲	cpp	▲
◆		DeepSeek-Coder-33b-instruct		47.17		80.02		52.03		65.13		62.36	
◆		DeepSeek-Coder-7b-instruct		45.92		80.22		53.34		65.8		59.66	
◆		OpenCodeInterpreter-DS-6.7B		45.42		73.2		51.41		63.85		60.01	
◆		Phind-CodeLlama-34B-v2		44.5		71.95		54.06		65.34		59.59	
◆		Phind-CodeLlama-34B-v1		43.42		65.85		49.47		64.45		57.81	
◆		Phind-CodeLlama-34B-Python-v1		41.88		70.22		48.72		66.24		55.34	
◆		CodeLlama-70b-Instruct		39.83		75.6		47.2		57.76		48.45	
◆		WizardCoder-Python-34B-V1.0		39.5		70.73		44.94		55.28		47.2	
●		CodeLlama-70b		39.33		52.44		44.72		56.52		49.69	
●		DeepSeek-Coder-33b-base		39.33		52.45		43.77		51.28		51.22	
●		CodeLlama-70b-Python		38.75		55.49		45.96		56.52		49.69	
●		StarCoder2-15B		37		44.15		33.86		44.24		41.44	
●		DeepSeek-Coder-7b-base		35.5		45.83		37.72		45.9		45.53	

What you need to train (code) LLMs from scratch



Performance scalability



Data scalability



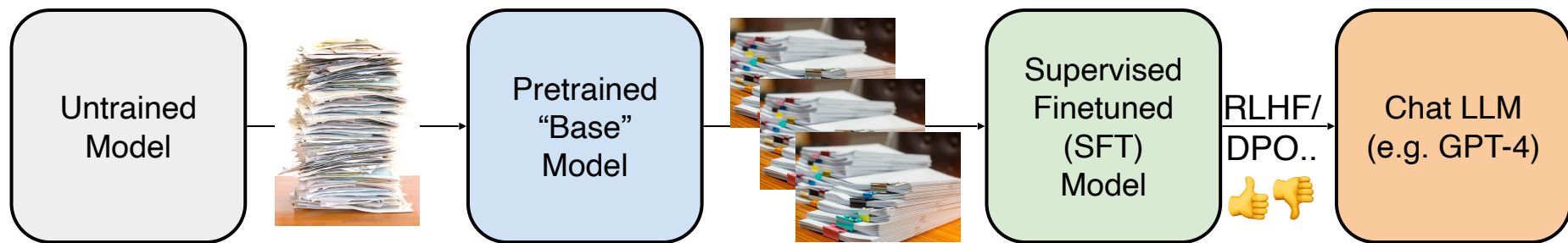
Hardware scalability

From GPT 1 → 4

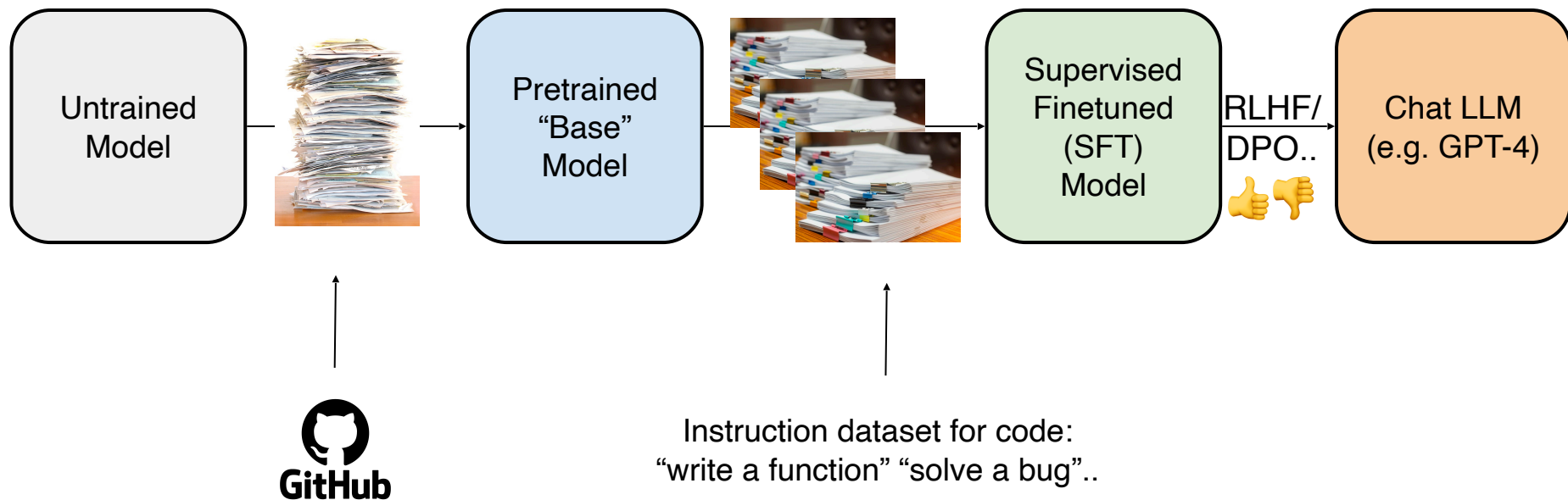
	Dataset size (Billion tokens)	Model size (Billion parameter)	
GPT 1:	1-2	0.11	} 100x
GPT 2:	10-20	1.4	
GPT 3:	300	175	} 2000x
GPT 4:	10'000	1'800	

↙ **GPT-4 cost: ~\$100M**

Training Generative AI Models



Training Code LLMs



The Landscape of base open code LLMs



- CodeLlama
- CodeLlama-Instruct
- 7B, 13B, 70B



BigCode

- The Stack dataset
- StarCoder & StarCoder2
- 3B, 7B, 15B sizes
- StarChat2 (with H4 team)



deepseek coder

- DeepSeek-Coder
- DeepSeek-Coder-Instruct
- 1B, 7B, 33B

Others: CodeQwen from Qwen team, CodeGen from Salesforce, StableCode from StabilityAI...

The gradient of model releases

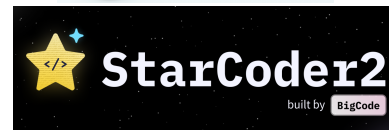
closed model APIs
model weights not available



open model weights
no access to training data or code



fully open model
full access to model/code/data



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



Open & Responsible Research on LLMs

Open-access datasets

[Data inspection](#)

[Opt-out available](#)

[PII removal](#)

[Attribution](#)

Open-access models

[Model weights available](#)

[Fine-tuning scripts](#)

[Low-precision inference](#)

Reproducible research

[Data preprocessing scripts](#)

[Model training framework](#)

[R&D notebooks](#)

[Evaluation Harness](#)

Documentation

[Dataset cards](#)

[Model cards](#)

[Governance card](#)

[Intellectual property](#)

[Code of conduct](#)

[OpenRAIL licenses](#)

From SantaCoder to StarCoder2 🚀



SantaCoder
Dec 2022

1.1B code generation model
3 languages
18% Python score
Transparent dataset
Open Access



StarCoder
May 2023

15B code generation model
80+ languages
33% Python score
Transparent dataset
Open Access

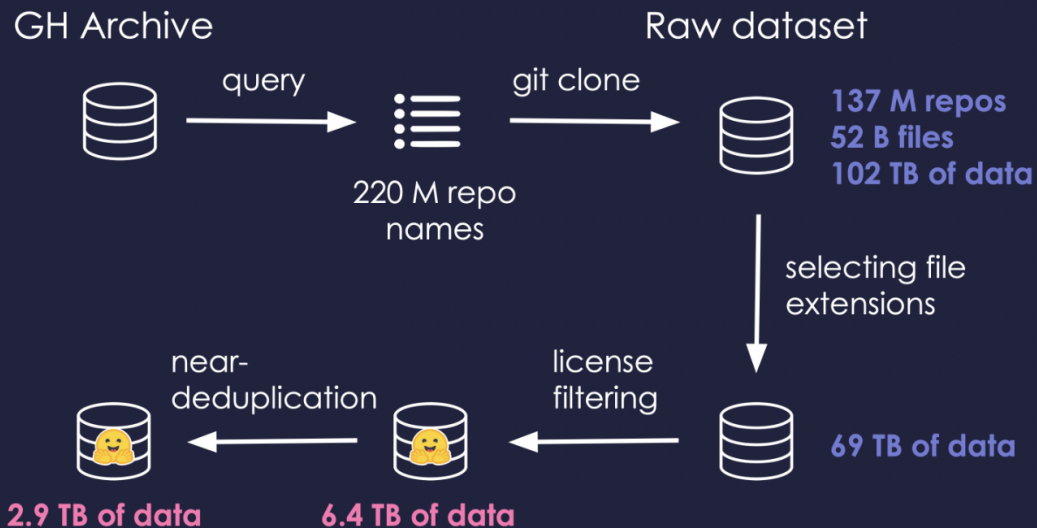


StarCoder2
Feb 2024

15B code generation model
600+ languages
46% Python score
Transparent dataset
Open Access



The Stack: data collection

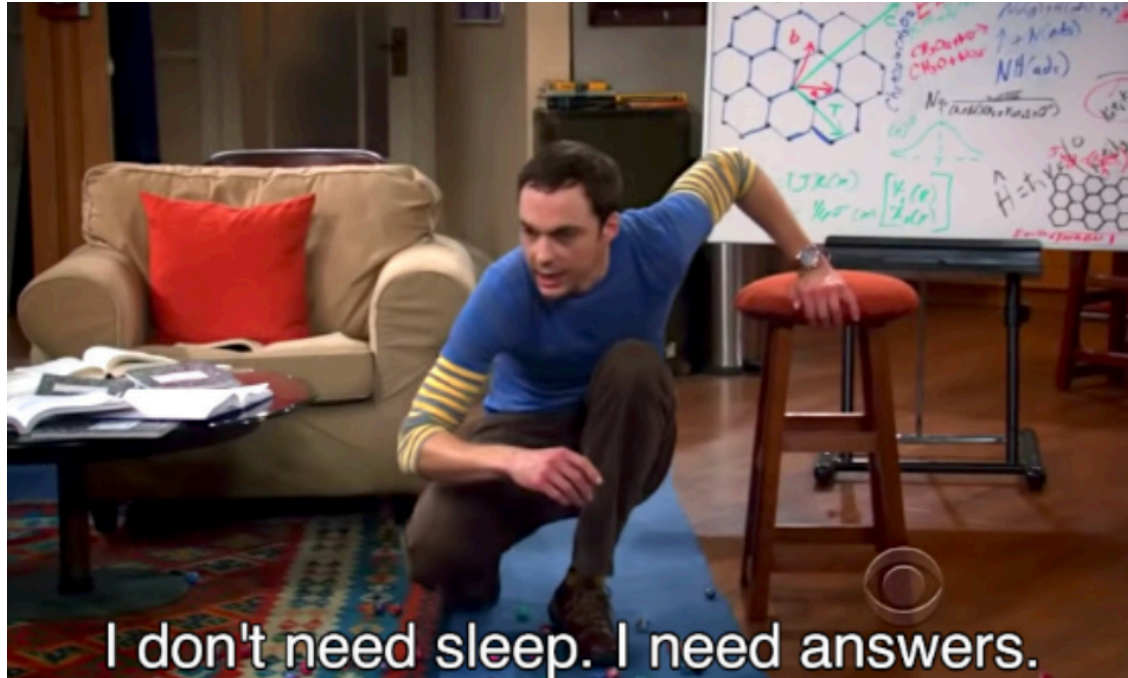


Find the filtered and deduplicated datasets at: www.hf.co/bigcode

StarCoderData

800 GB of code in 86 programming languages, with GitHub Issues, Jupyter Notebooks and Git Commits

where did the 6TB go?



Data filtering

- Near-deduplication
- Language selection & quality inspection
- Decontamination
- Personal Identifiable Information (PII) removal

StarCoder

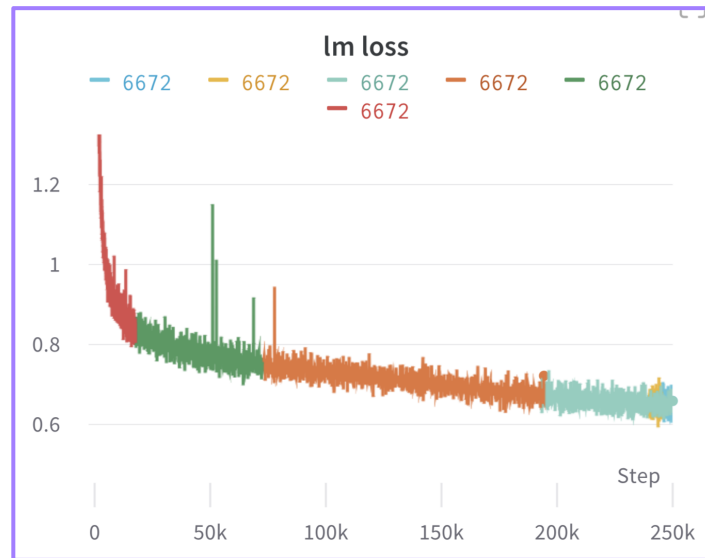
Model size: 15B parameters

Context length: 8096 tokens

Infrastructure: 512 A100 GPUs

Training length: 1T tokens / 250k steps

Training time: 24 days



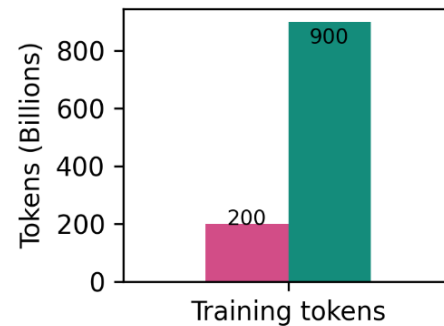
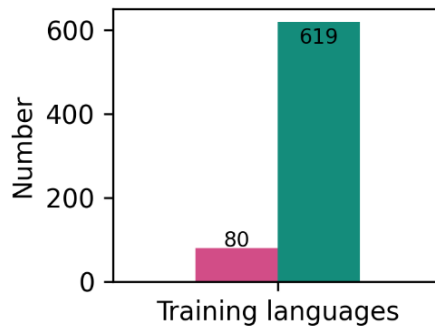
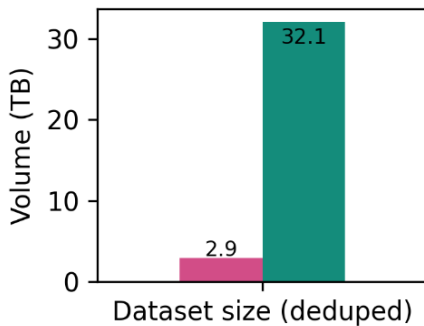
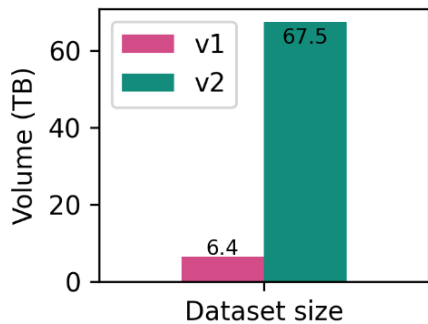
Best open LLM for code at the time of release!



The Stack v2

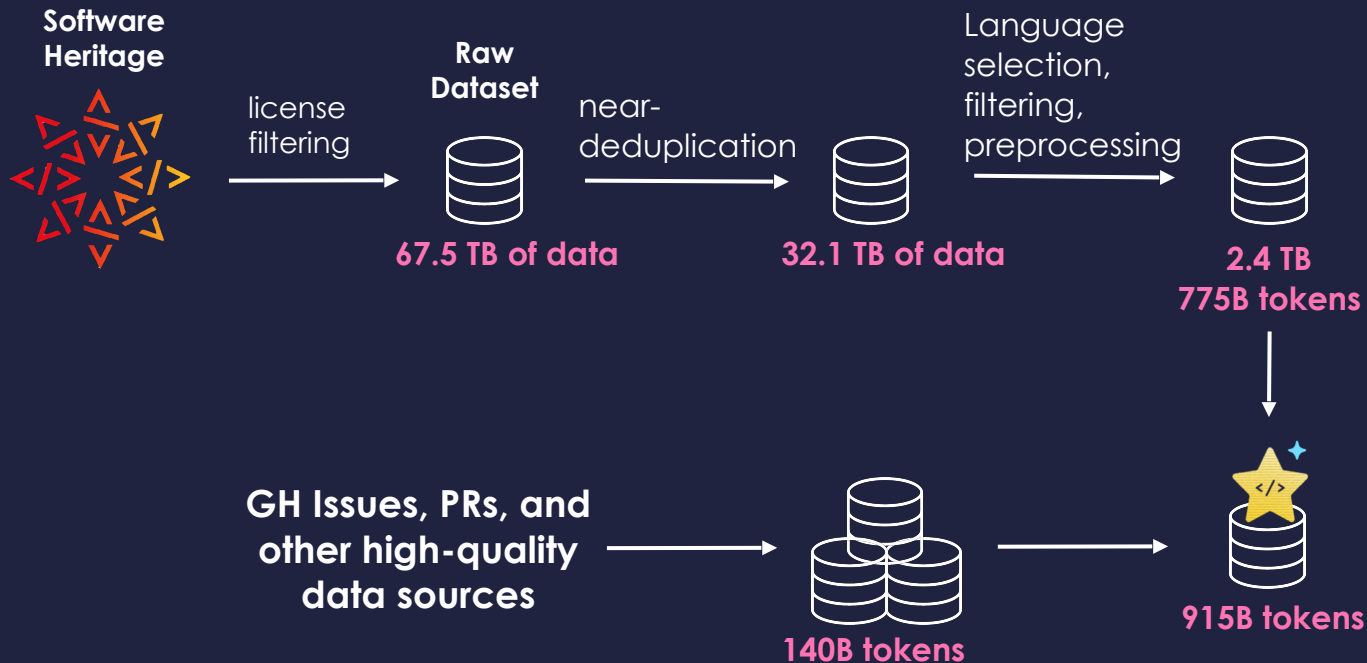


×



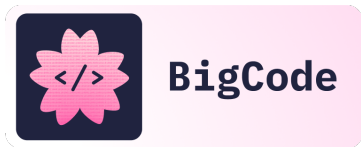
The Stack v2

Data collection



Extra sources

- Jupyter notebooks: structured (code & markdown pairs) vs scripts
- Kaggle notebooks
- GitHub issues and pull requests
- LHQ
- Wikipedia, Arxiv, OpenWebMath



The Stack: data inspection + opt-out



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.

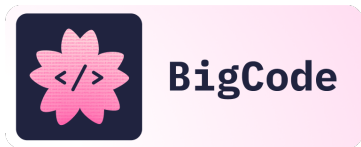
This tool lets you check if a repository under a given username is part of The Stack dataset. Would you like to have your data removed from future versions of The Stack? You can opt-out following the instructions [here](#).

The Stack version:

v1.1

Your GitHub username:

Check!



The Stack: data inspection + opt-out

bigcode-project / opt-out-v2

Code Issues 1.8k Pull requests Discussions Actions Projects Security

Opt-out request for nuprl #54

Closed arjunguha opened this issue on Nov 7, 2023 · 2 comments

arjunguha commented on Nov 7, 2023

I request that the following data is removed from The Stack and StackOverflow:

- nuprl/TypeWeaver

Note: If you don't want all resources to be included just remove the elements from the list above. If you would like to exclude all repositories and resources just add a single element "all" to the list.

arjunguha commented on Nov 7, 2023

This is a benchmark, and was used in the StarCoder paper. So best not to train on it. :)

lvwerra closed this as completed 2 days ago

lvwerra commented 2 days ago

Your opt-out request has been processed and your data was removed in version v2.0.1 of The Stack and all future versions. Also your data was not used for the training of StarCoder2.

[PROCESSED]

StarCoder2

Model size: 15B, 7B, 3B

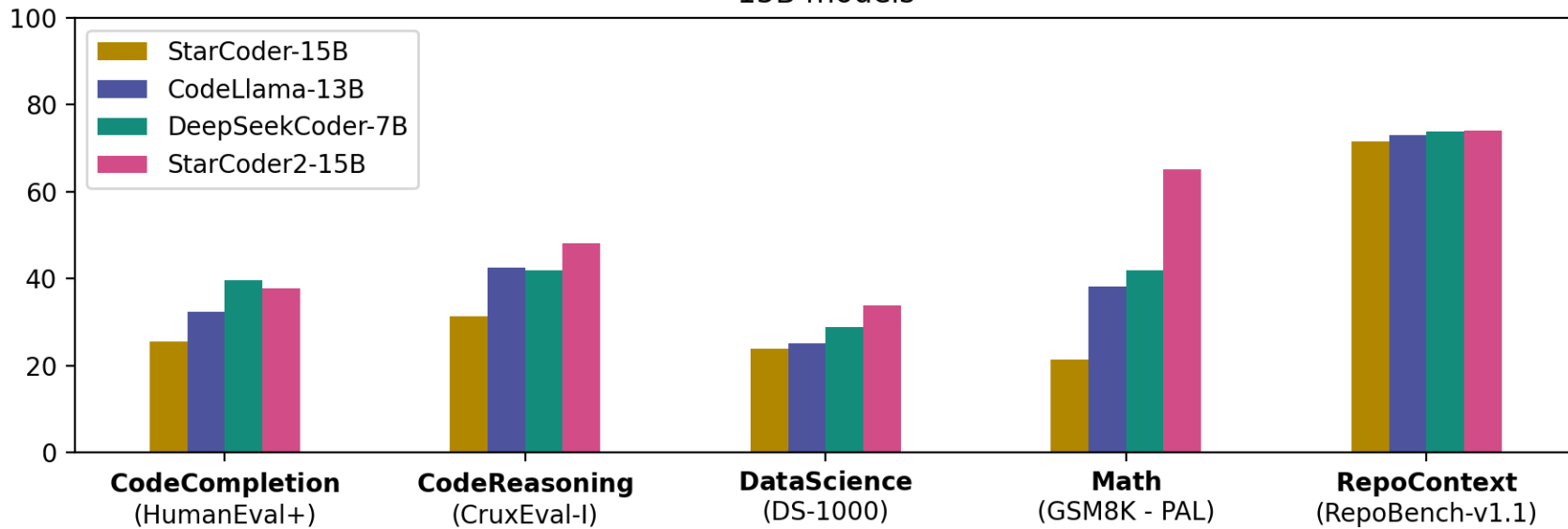
Context length: 16k tokens

Supports repository level context

Trained on 4T+ tokens

StarCoder2

15B models



Tooling

Auto-complete


```
Users > swayam > Desktop > Python main.py > ...
1 def is_prime(num):
2   return False

def is_prime(num):
    if num == 2:
        return True
    if num % 2 == 0:
        return False
    for i in range(3, num, 2):
        if num % i == 0:
            return False
```

<https://marketplace.visualstudio.com/items?itemName=HuggingFace.huggingface-vscode>

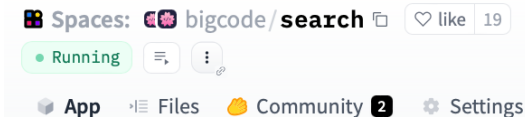
Membership test

```
Users > swayam > Desktop > Python main.py > is_prime
1 def is_prime(num):
2   return False
3
4 def is_prime(num):
5   if num == 2:
6     return True
7   if num % 2 == 0:
8     return False
9   for i in range(3, num, 2):
10    if num % i == 0:
11    return False
```

Highlighted code was found in the stack. 

Source: HF Code Autocomplete (Extension) [Go to stack search](#)

Dataset Search



StarCoder: Dataset Search 🔍

When using [StarCoder](#) to generate code, it might produce close or exact copies of code in the pretraining dataset. Identifying such cases can provide important context, and help credit the original developer of the code. With this search tool, our aim is to help in identifying if the code belongs to an existing repository. For exact matches, enclose your query in double quotes.

This first iteration of the search tool truncates queries down to 200 characters, so as not to overwhelm the server it is currently running on.

Query

<https://huggingface.co/spaces/bigcode/search>

Source: [MrGlockenspiel/Q_rsqrt-in-](#)

[Rust/src/main.rs](#)

Language: rust

License: WTFPL

```
<reponame>MrGlockenspiel/Q_rsqrt-in-Rust
use std::io::{self, BufRead};
use std::mem;
use std::time::Instant;

fn q_rsqrt(number: f32) -> f32 {
    let mut i: i32;
    let x2: f32;
    let mut y: f32;
    const THREEHALVES: f32 = 1.5;

    x2 = number * 0.5;
    y = number;

    // Evil floating point bit level hacking
    i = unsafe { mem::transmute(y) };

    // What the fuck?
    i = 0x5f3759df - (i >> 1);
    y = unsafe { mem::transmute(i) };

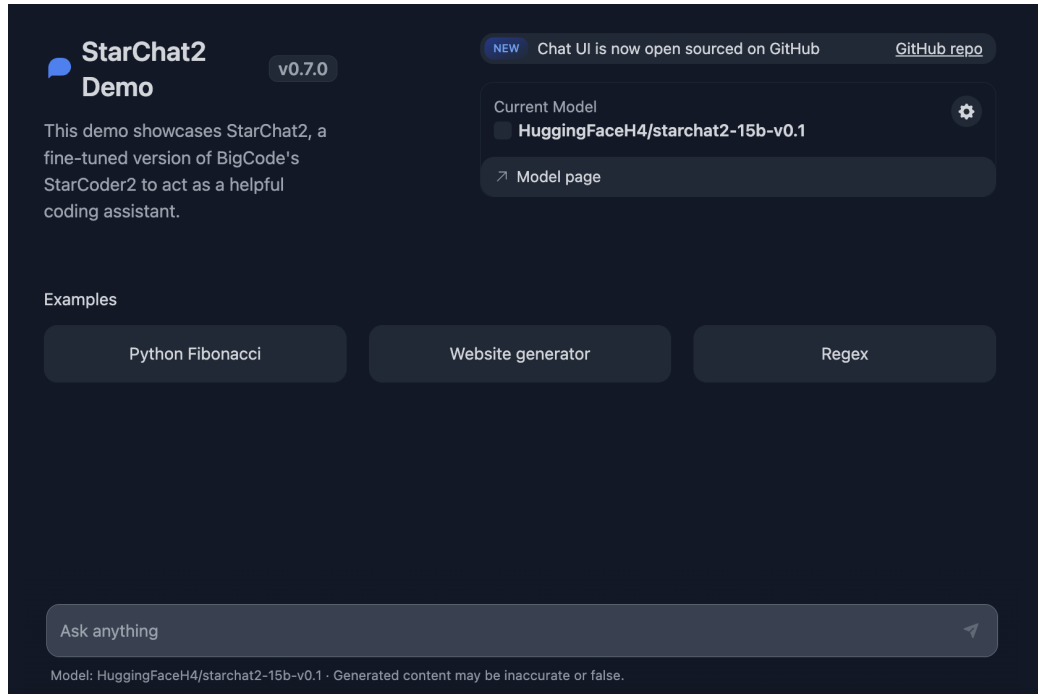
    // 1st iteration
    y = y * (THREEHALVES - (x2 * y * y));

    // 2nd iteration, this can be removed
    // y = y * (THREEHALVES - (x2 * y * y));

    return y;
}
```


Customize Code Models: Chat assistant

Instruction-tune a code model: Mix different open chat and code datasets <https://hf.co/spaces/HuggingFaceH4/starchat2-playground>



The screenshot displays the StarChat2 Demo interface. At the top left, it says "StarChat2 Demo" with a version tag "v0.7.0". A notification banner at the top right states "NEW Chat UI is now open sourced on GitHub" with a "GitHub repo" link. The main text describes the demo as showcasing StarChat2, a fine-tuned version of BigCode's StarCoder2, acting as a helpful coding assistant. Below this, there is a "Current Model" section showing "HuggingFaceH4/starchat2-15b-v0.1" with a settings gear icon and a "Model page" link. Under the "Examples" section, there are three buttons: "Python Fibonacci", "Website generator", and "Regex". At the bottom, there is a text input field with the placeholder "Ask anything" and a send arrow. A footer note reads: "Model: HuggingFaceH4/starchat2-15b-v0.1 · Generated content may be inaccurate or false."

Customize Code Models: Code completion

- Fine-tune an open code model on your codebase: <https://huggingface.co/blog/personal-copilot>

Personal Copilot: Train Your Own Coding Assistant

Published October 27, 2023

[Update on GitHub](#)



[smangrul](#)
[Sourab Mangrulkar](#)



[sayakpaul](#)
[Sayak Paul](#)

In the ever-evolving landscape of programming and software development, the quest for efficiency and productivity has led to remarkable innovations. One such innovation is the emergence of code generation models such as [Codex](#), [StarCoder](#) and [Code Llama](#). These models have demonstrated remarkable capabilities in generating human-like code snippets, thereby showing immense potential as coding assistants.

However, while these pre-trained models can perform impressively across a range of tasks, there's an exciting possibility lying just beyond the horizon: the ability to tailor a code generation model to your specific needs. Think of personalized coding assistants which could be leveraged at an enterprise scale.

Leaderboards

The screenshot shows a web browser window displaying the 'Big Code Models Leaderboard' on HuggingFace Spaces. The page title is 'Big Code Models Leaderboard' and the URL is 'huggingface.co/spaces/bigcode/bigcode-models-leaderboard'. The page features a navigation bar with 'Spaces', 'bigcode', 'bigcode-models-leaderboard', 'like 706', 'Running', and 'Logs'. Below the navigation bar is the main title 'Big Code Models Leaderboard' with a star icon. A paragraph of text explains the leaderboard's purpose: 'Inspired from the Open LLM Leaderboard and Open LLM-Perf Leaderboard, we compare performance of base multilingual code generation models on HumanEval benchmark and MultiPL-E. We also measure throughput and provide information about the models. We only compare open pre-trained multilingual code models, that people can start from as base models for their trainings.'

Below the text is a navigation bar with 'Evaluation table', 'Performance Plot', 'About', and 'Submit results'. A search bar is present with the placeholder 'Search for your model and press ENTER...'. To the right of the search bar are filter buttons for 'all', 'base', 'instruction-tuned', and 'EXT external-evaluation'. The 'base' filter is currently selected.

The main content is a table with the following columns: 'T', 'Model', 'Win Rate', 'humaneval-python', 'java', 'javascript', and 'cpp'. The table lists 13 models with their respective performance metrics. The 'EXT' column contains colored diamond icons (orange, green, or red) indicating the model's status.

T	Model	Win Rate	humaneval-python	java	javascript	cpp
EXT	OpenCodeInterpreter-DS-33B	54.67	75.23	54.8	69.06	64.47
	CodeQwen1.5-7B-Chat	53.83	87.2	61.04	70.31	67.85
EXT	CodeFuse-DeepSeek-33b	53.08	76.83	60.76	66.46	65.22
EXT	DeepSeek-Coder-33b-instruct	51.33	80.02	52.03	65.13	62.36
EXT	DeepSeek-Coder-7b-instruct	49.92	80.22	53.34	65.8	59.66
EXT	OpenCodeInterpreter-DS-6.7B	49.33	73.2	51.41	63.85	60.01
	Phind-CodeLlama-34B-v2	48.5	71.95	54.06	65.34	59.59
	Phind-CodeLlama-34B-v1	47.38	65.85	49.47	64.45	57.81
	Phind-CodeLlama-34B-Python-v1	45.81	70.22	48.72	66.24	55.34
	CodeQwen1.5-7B	44.92	50.79	42.15	50.07	48.35
	CodeLlama-70b-Instruct	43.33	75.6	47.2	57.76	48.45

Leaderboards

🏆 EvalPlus Leaderboard 🏆

EvalPlus evaluates AI Coders with rigorous tests.



<EvalPlus Tests>

#	Model	pass@1
1	🏆 GPT-4-Turbo (Nov 2023) 📈	≤81.7
2	🏆 GPT-4 (May 2023) 📈	≤79.3
3	🏆 claude-3-opus (Mar 2024) 📈	≤76.8
4	DeepSeek-Coder-33B-instruct 📈	≤75
5	OpenCodeInterpreter-DS-33B 📈❤️	≤73.8
6	WizardCoder-33B-V1.1 📈	≤73.2
7	OpenCodeInterpreter-DS-6.7B 📈❤️	≤72
8	speechless-codellama-34B-v2.0 📈❤️	≤71.3
9	GPT-3.5-Turbo (Nov 2023) 📈	≤70.7
10	Magocoder-S-DS-6.7B 📈❤️	≤70.7

Base Tests

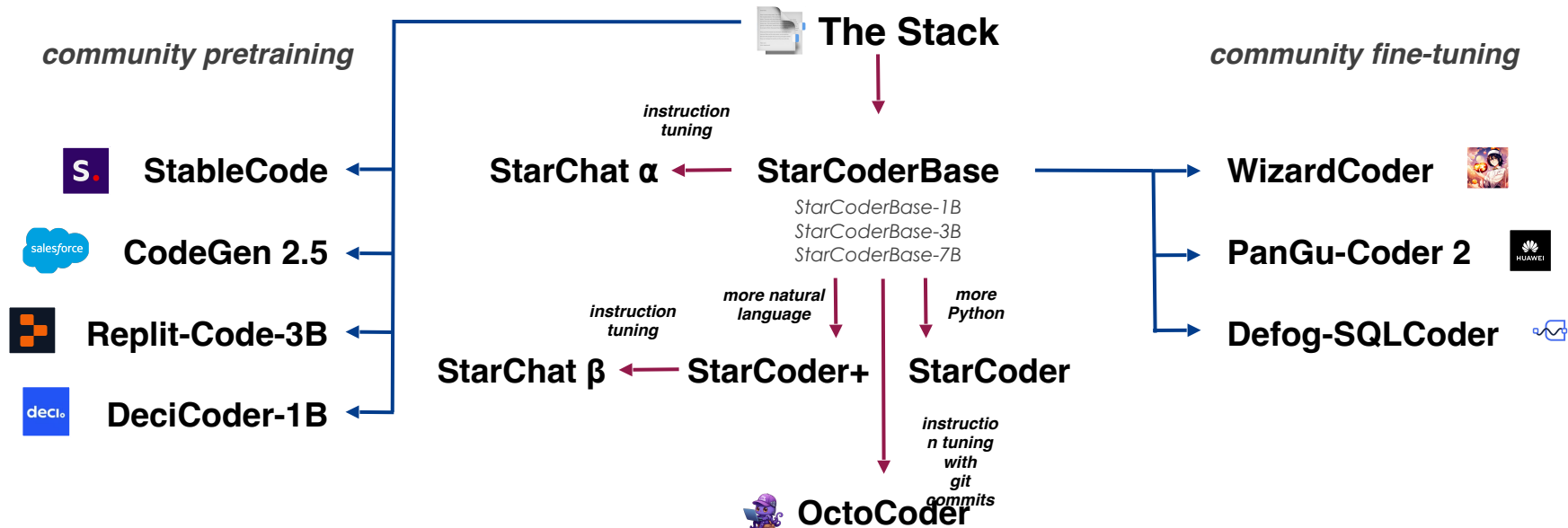
#	Model	pass@1
1	🏆 GPT-4 (May 2023) 📈	88.4
2	🏆 GPT-4-Turbo (Nov 2023) 📈	85.4
3	🏆 claude-3-opus (Mar 2024) 📈	82.9
4	DeepSeek-Coder-33B-instruct 📈	81.1
5	WizardCoder-33B-V1.1 📈	79.9
6	OpenCodeInterpreter-DS-33B 📈❤️	79.3
7	OpenCodeInterpreter-DS-6.7B 📈❤️	77.4
8	speechless-codellama-34B-v2.0 📈❤️	77.4
9	GPT-3.5-Turbo (Nov 2023) 📈	76.8
10	Magocoder-S-DS-6.7B 📈❤️	76.8

Leaderboards

The screenshot shows a web browser displaying the LiveCodeBench leaderboard. The page title is "LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code". Below the title are navigation buttons for "Paper", "Code", "Data", and "Home". A secondary navigation bar includes "Code Generation" (selected), "Self Repair", "Test Output Prediction", and "Code Execution". A date range selector shows "238 problems selected in the current time window" with dates from "01/09/2023" to "01/04/2024". A table lists the top 5 models with their ranks and performance metrics.

Rank	Model	Pass@1 ↓	Easy-Pass@1	Medium-Pass@1
1	GPT-4-Turbo-2024-04-09	44.1	81.4	33.7
2	GPT-4-Turbo-1106	39.6	81.9	24.5
3	GPT-4-0613	35.5	74.6	21.3
4	Claude-3-Opus	34.8	76.5	14.7
5	Gemini-Pro-1.5 (n=1)	28.6	56.5	17.3

BigCode Ecosystem



Challenges of a fully open collaboration

- **decision making**
 - decentralized decision making is more difficult
- **public scrutiny**
 - everybody can check code and datasets and report issues
- **maintenance**
 - public code base and datasets need to be kept up to date (e.g. opt-outs)
- **public timelines**
 - other projects can adapt their timeline to yours but not vice-versa

Future Directions

- **High quality datasets** for high and low resource languages
- **More data transparency and governance**
- **Evaluation** benchmarks & leaderboards
- **Smaller specialized models**

Thank you!

Contact: loubna@huggingface.co