Building Large Language Models For Code

KubeCon | CloudNativeCon

OPEN SOURCE SUMMIT

China 2023

Loubna Ben Allal | ML Engineer @ Hugging Face | 𝕏 LoubnaBenAllal1

Let's start with some Context 📖

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

```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, value, currency).
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2016-01-02 -34.01 USD
        2016-01-03 2.59 DKK
        2016-01-03 -2.72 EUR
    """
    expenses = []
    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split(" ")
        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
                         float(value),
                         currency))
    return expenses
```

Copilot

Replay

From GitHub Copilot to open Code Models 🚀

CodeGPT, CodeBERT, PLBART… → Codex, AlphaCode, CodeWhisperer

From GitHub Copilot to open Code Models 🚀

CodeGPT, CodeBERT, PLBART… → Codex, AlphaCode, CodeWhisperer → CodeParrot, PolyCoder, InCoder, CodeGen

Closed Models 😢

From GitHub Copilot to open Code Models 🚀

CodeParrot, PolyCoder, InCoder, CodeGen

Open questions: Performance, Transparency about training data, Multilinguality, Evaluation, User experience …

Hugging Face: From CodeParrot to StarCoder 🚀

CodeParrot

- 1.5B code generation model
- Python only
- **4%** Python score
- **Permissive data**
- **Open Access**

StarCoder

- 15B code generation model
- 80+ languages
- **33%** Python score: beats code-cushman-001 (Codex)
- **Permissive data**
- **Open Access**

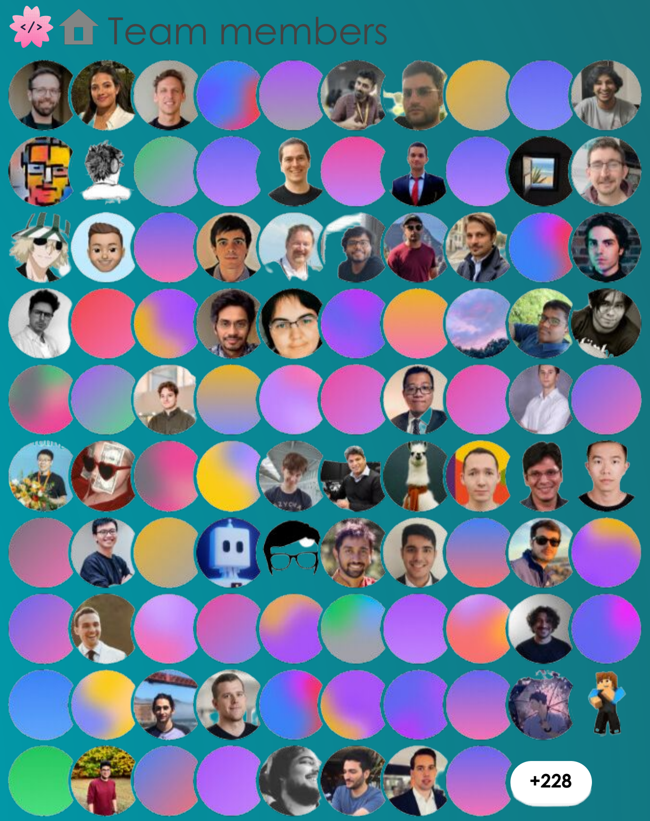
Today's talk

- The BigCode Community
- Dataset
- Architecture
- Model deployment

BigCode: open-scientific collaboration

We are building LLMs for code in a collaborative way:

- **500+** participants
- **30+** countries



Closed LLM development

- Training data and sources not disclosed
- Model weights not public
- Sending data to external APIs
- Not reproducible

Open LLM development 🌅

- Public data with inspection and opt-out tools
- Model weights public for fine-tuning
- On-prem deployment
- Full documentation

BigCode

Training LLMs for Code from scratch

Hundreds of GPU-hours, terabytes of data

But not just that!

BigCode

StarCoder

Dataset: The Stack

Public dataset with 6.4TB of permissively licensed source code from GitHub in 358 programming languages with a data inspection tool and opt-out mechanism

Training Data Curation

- Language selection & quality inspection
  - 86 languages
  - GitHub issues, git commits & Jupyter notebooks
- Deduplication
- Decontamination
- Personal Identifiable Information (PII) removal

BigCode

How to run preprocessing on large datasets

- **Load datasets** from the Hub using multiprocessing
- **filter()** and **map()** to apply a transformation using multiprocessing
- **Batched mapping:** Dataset.map() in batch mode

```python
from datasets import Dataset

dataset = Dataset.from_dict({"a": [0, 1, 2]})
# new column with 6 elements: [0, 1, 2, 0, 1, 2]
dataset.map(lambda batch: {"b": batch["a"] * 2}, batched=True)
```

https://github.com/bigcode-project/bigcode-dataset/blob/main/preprocessing/filtering.py

BigCode

Training

Architecture choices

What do people want from a code model?

- Fast inference

  → 15B parameters with code optimizations

- Cheap generations

  → Multi-Query Attention for reduced memory footprint

- Long context

  → Flash Attention to scale to 8,192 tokens context

- Bi-directional context

  → Fill-in-the-middle training objective

BigCode

Training setup

Infrastructure: 512 GPUs

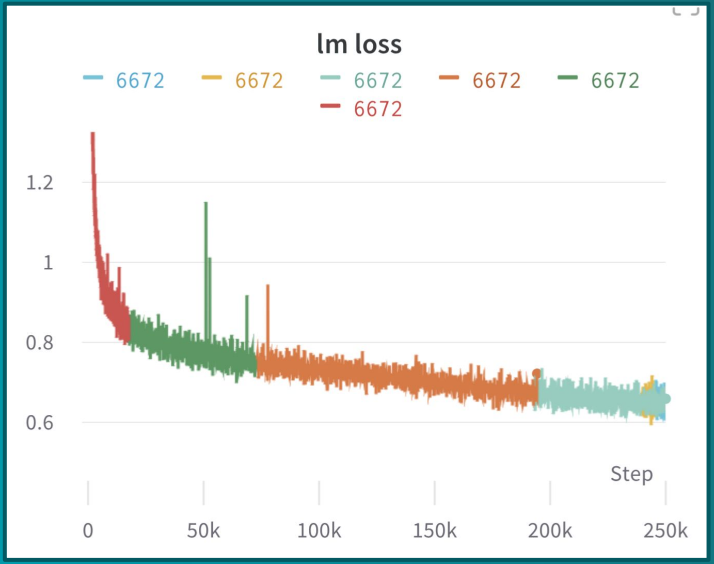
Model Distribution: TP=4, PP=4, DP=32

Batch size: 4M tokens
(or 512 at 8,192 sequence length)

Training length: 1T tokens / 250k steps

Training time: 24 days

Tool: Megatron-LM



"smooth sailing"

BigCode

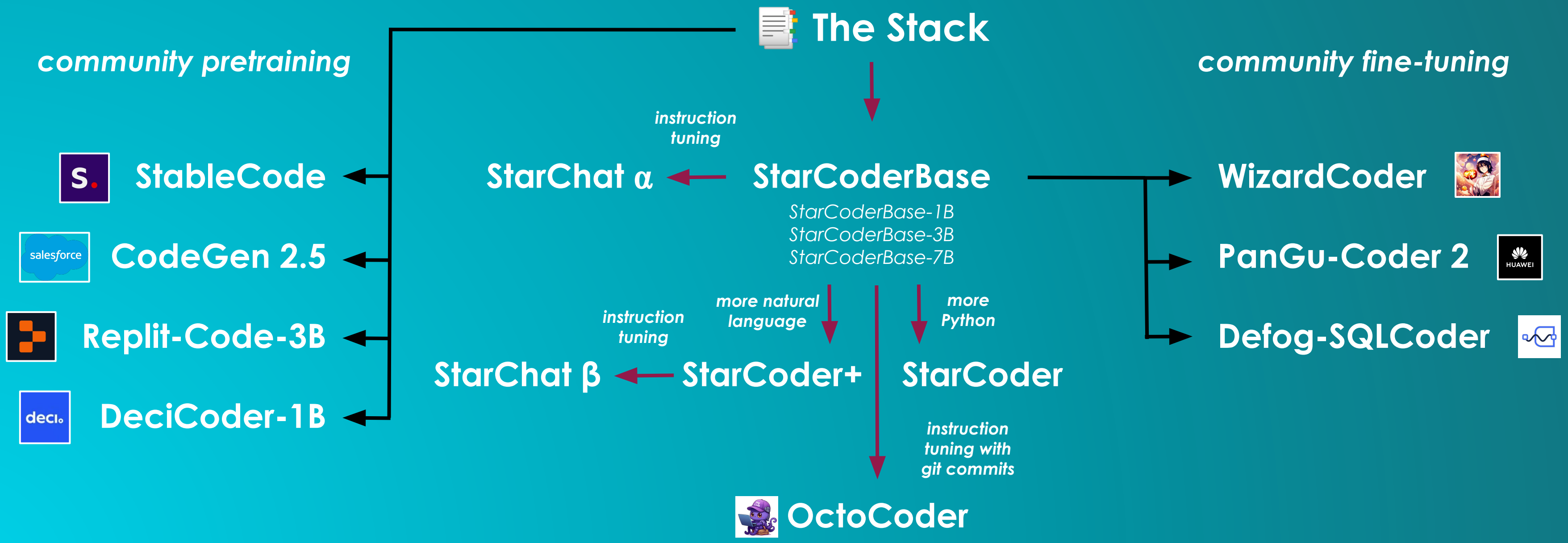
Fine-tuning models on low resources: PEFT 🤗

- Fine-tune a small number of (extra) parameters at **low computational cost** parameters with **comparable performance to full fine-tuning**

- Only push and load adapter weights for inference → **low storage cost**

https://github.com/bigcode-project/starcoder

BigCode Ecosystem



Deploying Large Language Models (for Code)

🤗 Hugging Face Inference endpoints

A Better Way to Go to Production

Scale your machine learning while keeping your costs low

Before

- Struggle with MLOps and building the right infrastructure for production.
- Wasted time deploying models slows down ML development.
- Deploying models in a compliant and secure way is difficult & time-consuming.
- 87% of data science projects never make it into production.

After

- Don't worry about infrastructure or MLOps, spend more time building models.
- A fully-managed solution for model inference accelerates your ML roadmap.
- Easily deploy your models in a secure and compliant environment.
- Seamless model deployment bridges the gap from research to production.

🤗 Text-generation-inference (TGI)

Tensor Parallelism

Token Streaming

Metrics and monitoring

Quantization

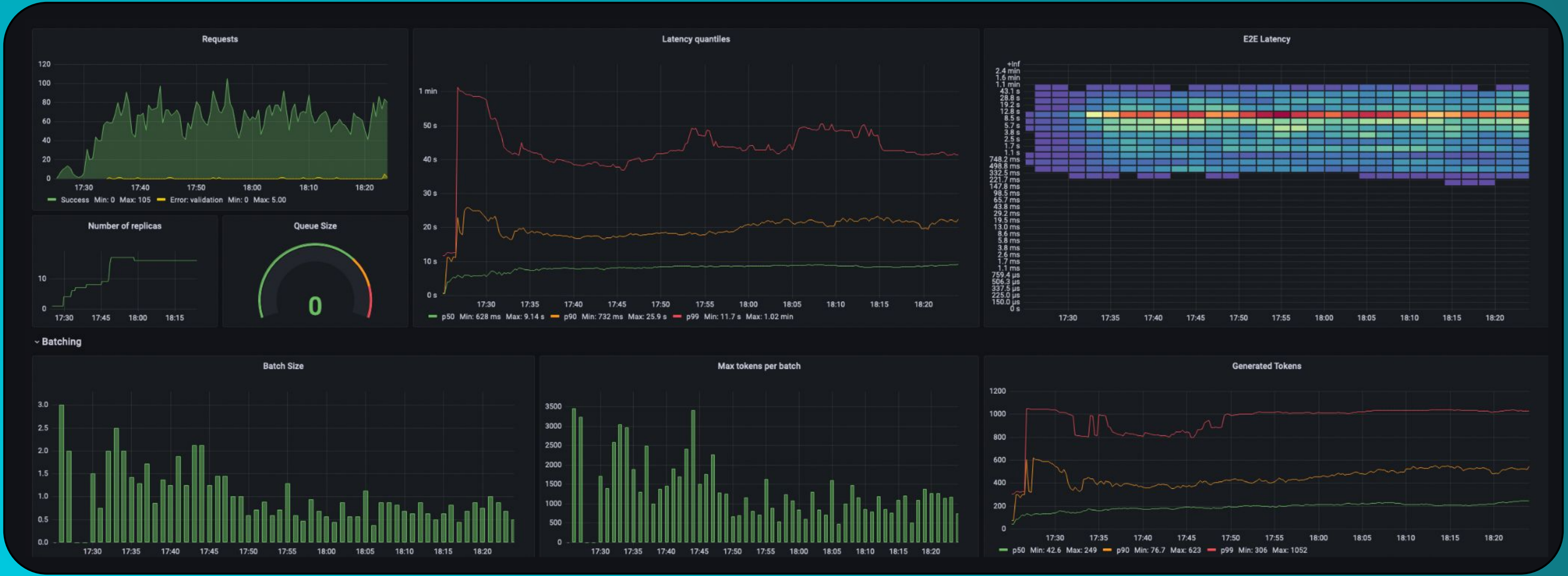
Optimizations

Security

TGI supports most popular LLMs, such as

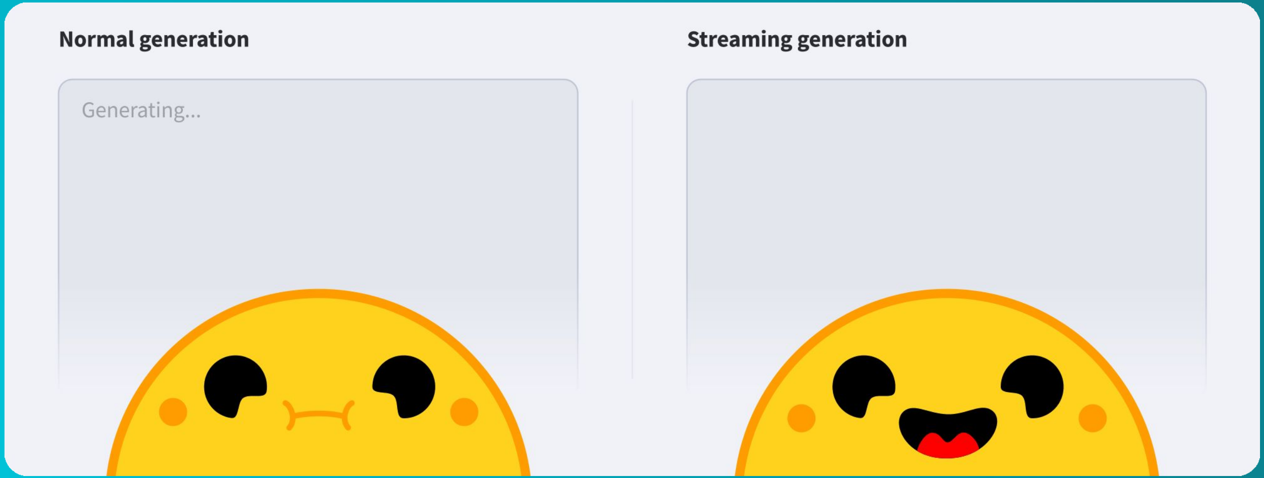
Falcon | StarCoder and SantaCoder | LLaMA, Galactica and OPT | GPT-NeoX

📊Production ready: Tracing mechanism & Warmup



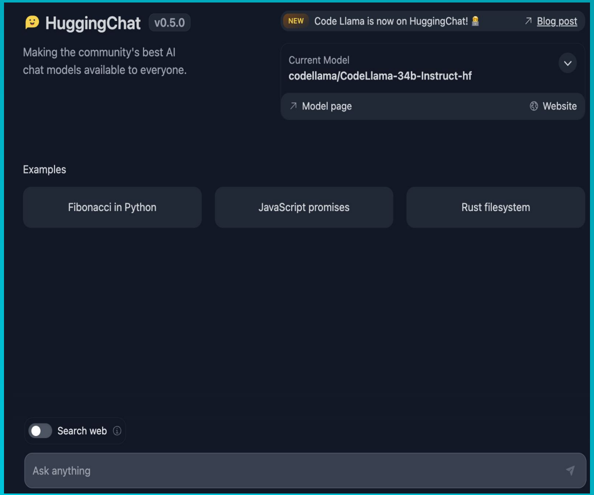
Optimizations & user experience

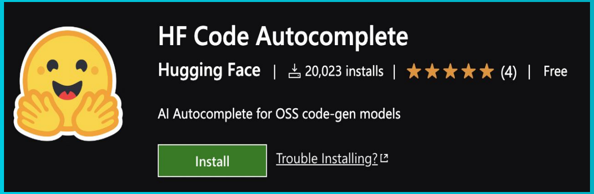
- Optimized for latency
- Continuous batching: for handling concurrent requests
- Token streaming: reduce perceived latency and improve interactivity



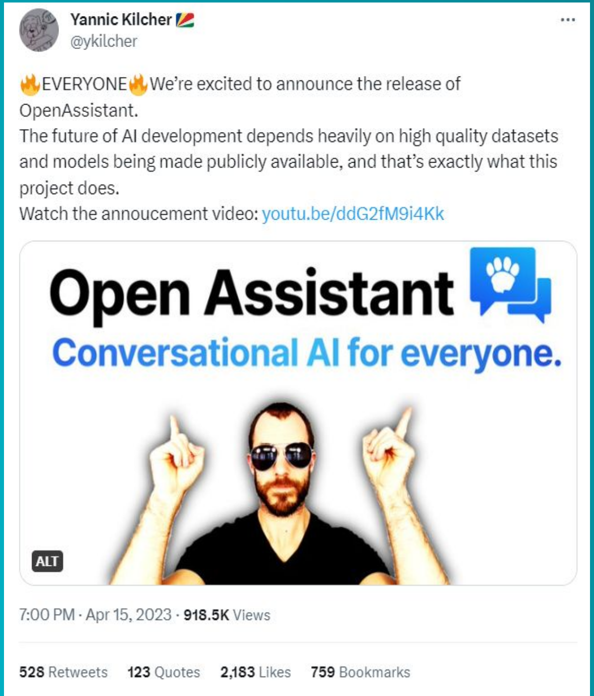
Some users

HuggingChat

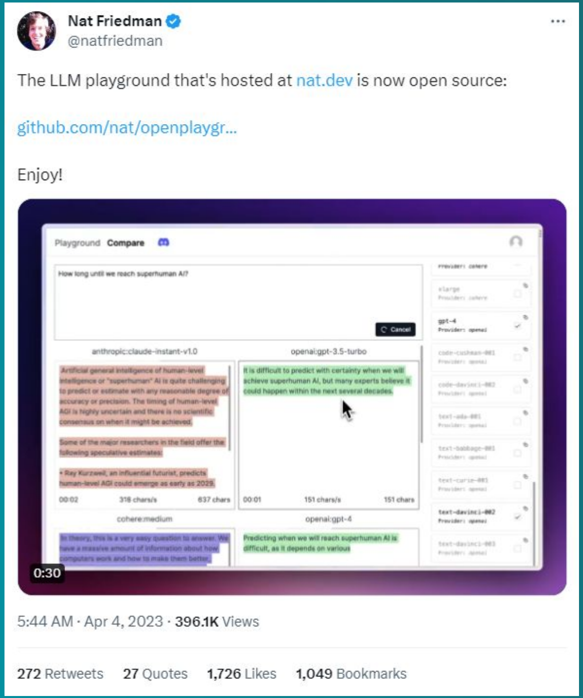




OpenAssistant



nat.dev



Kubernetes at Hugging Face

All Hugging Face Infrastructure uses Kubernetes

- **8** production clusters, **800** nodes
- Hub, API Endpoints, Dataset Server, Spaces...

Very dense clusters:

- Use of memory swap feature
- Up to **250 pods** in a node

Re-compilation of containerd to pull images faster

- 30% faster checksum operations
  https://go-review.googlesource.com/c/go/+/353402

Questions

KubeCon CloudNativeCon
OPEN SOURCE SUMMIT
China 2023

Hugging Face