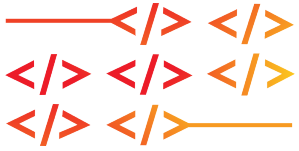


CodeCommons

Next generation infrastructure
for enabling transparent AI on code
and massive analysis of software source code

Inria | Software Heritage — THE GREAT LIBRARY OF SOURCE CODE | cea | TWEAG by Modus Create

Sant'Anna School of Advanced Studies – Pisa | Università degli Studi di Torino | Università di Pisa | Alma Mater Studiorum Università di Bologna | AboutCode

GENCI — HPC at the service of knowledge

bpifrance — SERVIR L'AVENIR

FRANCE 2030

Software Heritage and Generative AI, first contacts

October 19, 2023

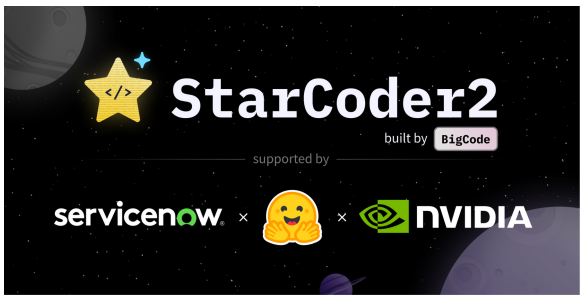
Software Heritage Statement on Large Language Models for Code

Principles

1. Knowledge derived from the Software Heritage archive must be given back to humanity, rather than monopolized for private gain. The resulting *machine learning models* must be made available under a suitable open license, together with the documentation and toolings needed to use them.
2. The *initial training data extracted from the Software Heritage archive* must be fully and precisely identified by, for example, publishing the corresponding SWHID identifiers (note that, in the context of Software Heritage, public availability of the *initial training data* is a given: anyone can obtain it from the archive). This will enable use cases such as: studying biases (fairness), verifying if a code of interest was present in the training data (transparency), and providing appropriate attribution when generated code bears resemblance to training data (credit), among others.
3. Mechanisms should be established, where possible, for authors to exclude their archived code from the training inputs before model training begins.

February 2024

Yes, it's possible!



But it's hard…

GENERATIVE AI FOR CODE : OPEN ISSUES

Gousios et al. GHTorrent
: GitHub's data from a firehose, MSR 2012

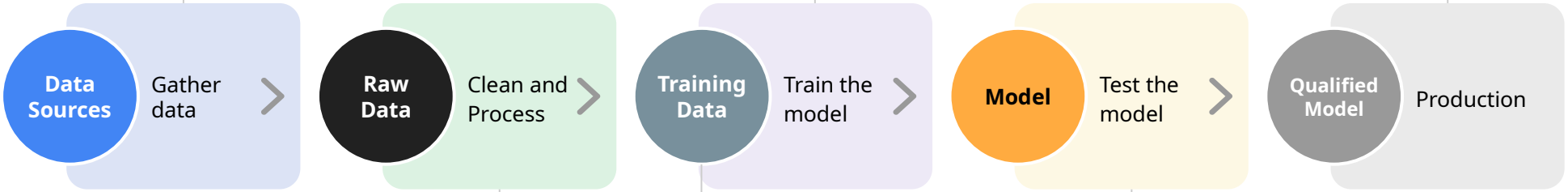
Collect source code, issues, PR, discussions, etc. **is very expensive**. **Redoing** it over and over again **is an anti-ecological waste**.

Lefeuvre et al. Fingerprinting and Building Large Reproducible Datasets REP'23

No precise identification and **lack of availability** of training data are huge obstacles to **transparency** and **reproducibility**.

Sallam et al.
ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns
2023

Lack of **traceability** of generative AI outputs make it **irrespective of authors**



Building a quality training set is a **very complex task,** redoing it over and over again behind closed doors is a waste of energy and human resources

Gunasekar et al. « Textbooks Are All You Need »
2023
https://arxiv.org/abs/2306.11644

Extracting qualified subsets for training is **difficult** and time consuming.

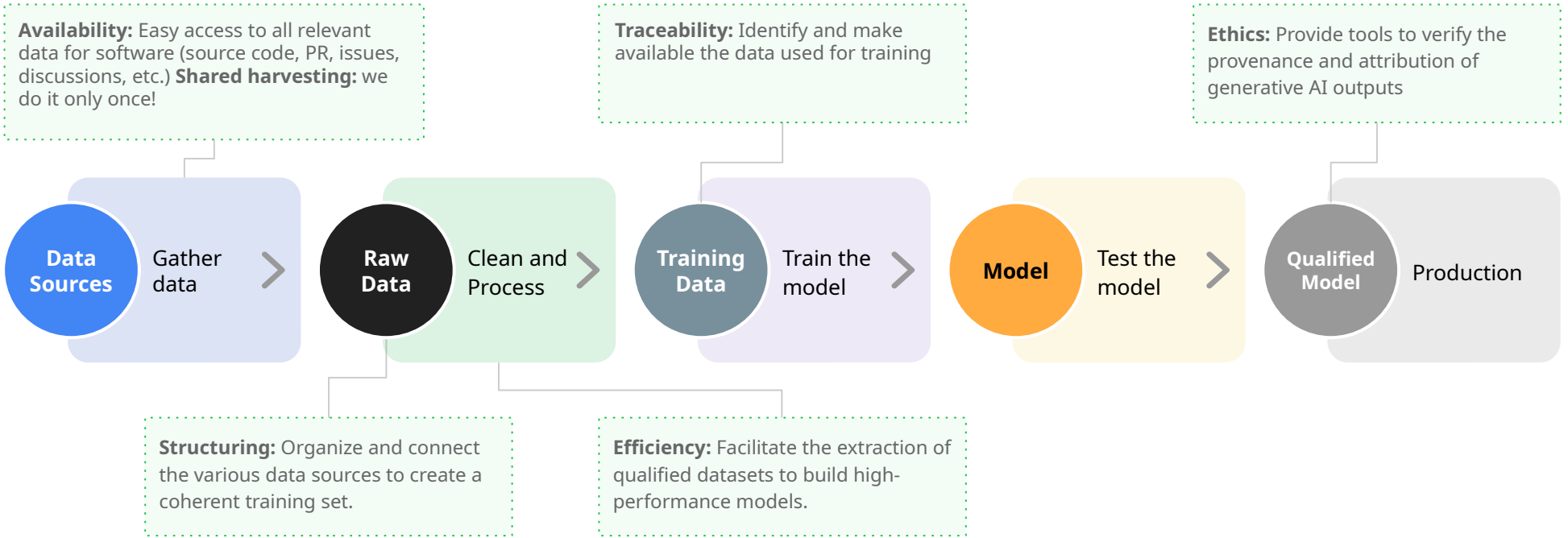
Ledivarec et al.
HyperDiff: Computing Source Code Diffs at Scale
ASE 2023

Extracting **quality subsets** should allow to **specialize** LLMs to perform **quality programming and software engineering tasks**.

Fan et al. Large language models for software engineering: Survey and open problems
FoSE 2023

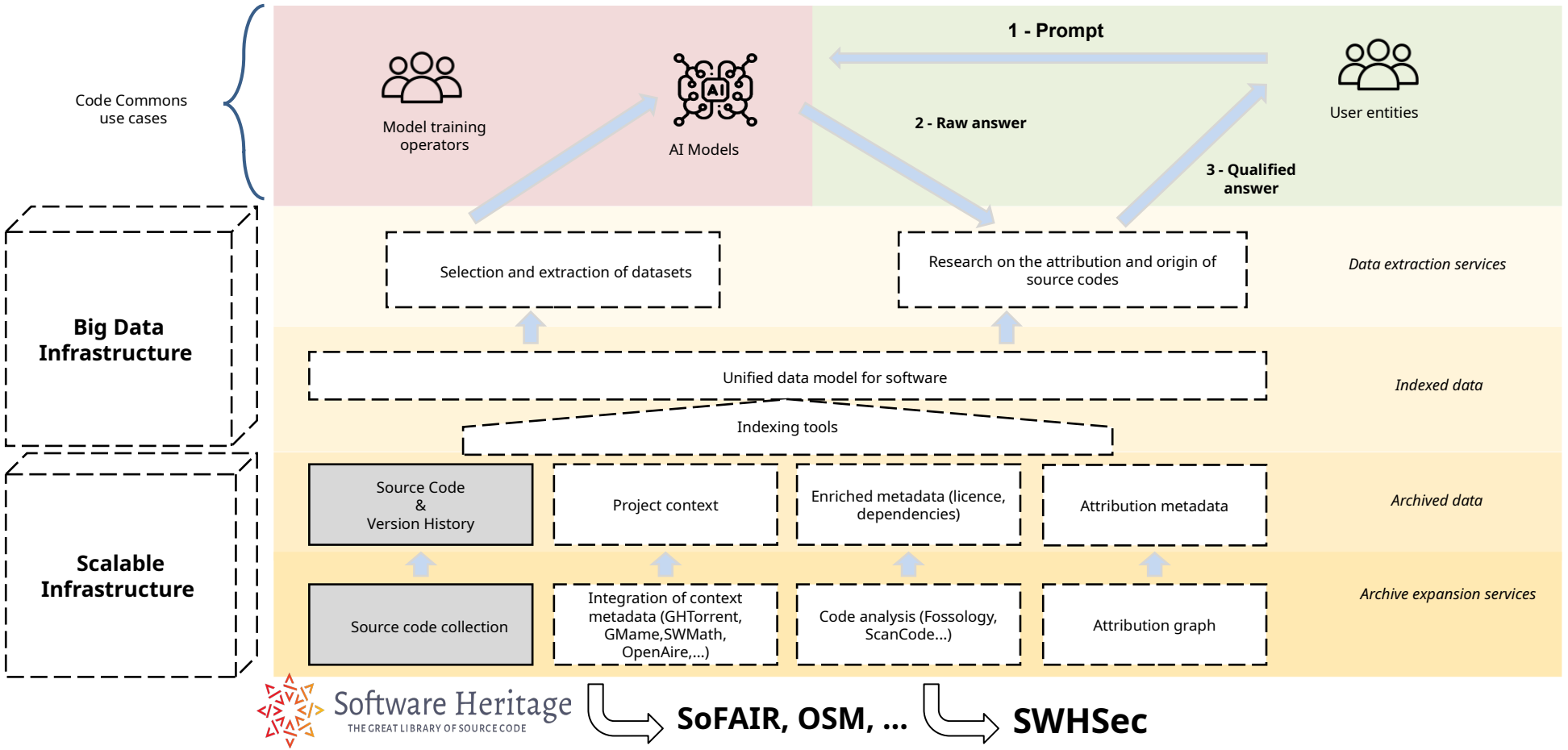
Issues

A STEP FORWARD: CodeCommons



Solutions

CodeCommons: bird's eye view (technical focus)



CodeCommons

Open, responsible, and transparent AI: Our shared goal

CodeCommons is an ambitious project to create the world's most comprehensive digital commons for code

Building on the existing foundation of Software Heritage, the largest publicly available source code archive, CodeCommons aims to bring into one place all the **critical** and **qualified** information needed to create **smaller**, **better** datasets for the next generation of AI tools.

At its core, the project prioritizes transparency and traceability, enabling model builders and users to **respect creators' rights** while promoting **sovereign** and **sustainable** AI.

Learn more

Meet the teams