

# Open Code LLMs and Software Heritage



Image created with DALL-E

Roberto Di Cosmo

Inria and Université Paris Cité

Director, Software Heritage

<https://dicosmo.org>

@rdicosmo

# Large Language Models for Code

---



Image created with DALL-E

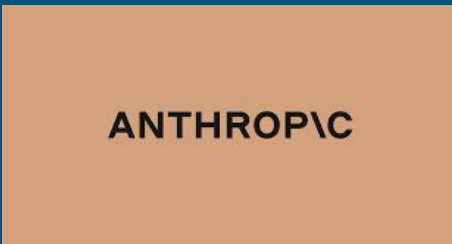
*Software source code is **massively used** for building Large Language Models.*

**Independently** of what we do, there is no turning back.

The **real question** is *how* they should be built and *whom* they should benefit.

*Let's have a candid look around  
US.*

# Closed model APIs



# 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

And more:

- limits user freedom  
(personal data leakage)

# 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

And more:

- limits user freedom  
(personal data leakage)

# Open model weights

## Training data is not disclosed

- Content creators don't know if their data is used
- There's no way to remove it
- Can't inspect data for biases
- Potential benchmark contamination

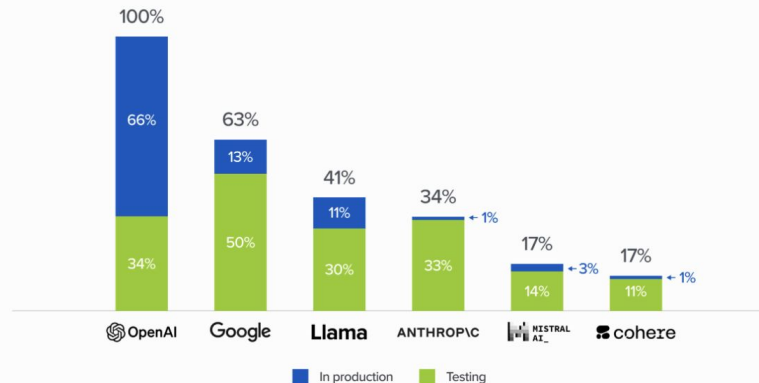
**This is not what “open” should mean.**

**Can we change all this? How?**

# A window of opportunity: market

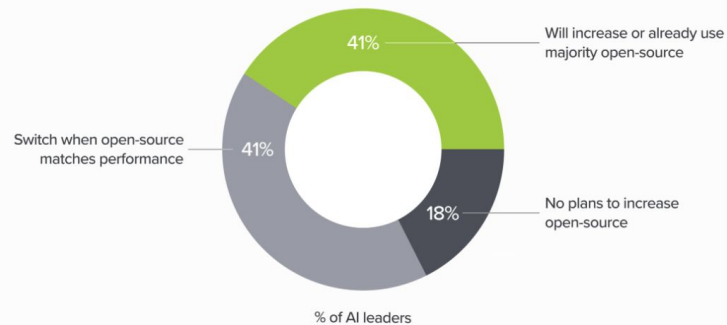
Which model providers are enterprises using?

a16z Growth



Enterprise expectations for open source usage in 2024 and onward

a16z Growth



<https://a16z.com/generative-ai-enterprise-2024/>

*LLMs follow a winners take all dynamics...*

*... but companies want "open source" models*

# A window of opportunity: regulations

## Open source AI definition

*Sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system.*



## AI Act

**Art. 53: Exception for providers of AI models released under a free and open-source licence[...]** and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available.



Software Heritage in this picture





## Largest archive of source code digital commons built since 2015

Cultural Heritage



Industry



Research

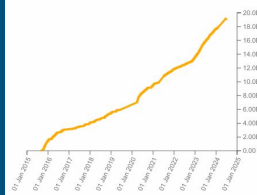


Public Administration



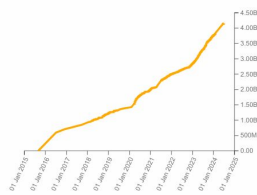
Source files

19,191,593,778



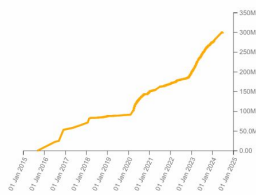
Commits

4,152,398,882



Projects

300,482,385



Directories

15,388,884,076

Authors

75,379,404

Releases

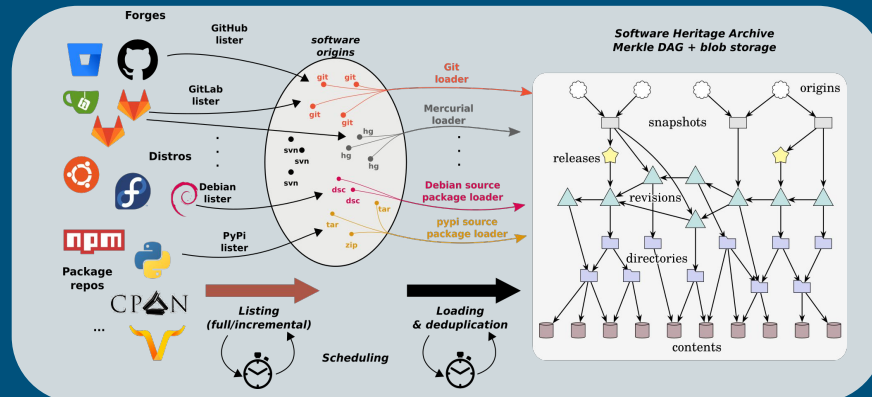
94,476,739

Ensures **availability**  
Guarantees **integrity**  
Enables **traceability**



of all source code

Unique dataset for machine learning,  
an infrastructure for transparency and accountability



500+ code hosting platforms

All versions, full development history  
In a single giant Merkle Graph

- $50 \times 10^9$  nodes
- $700 \times 10^9$  edges
- ~ 2 PB storage

# Example simple tasks

Find all contents that have as “popular filename” \*.v

```
select swhid, lower(trim(from_utf8(filename))) from
popular_content_name where
lower(trim(from_utf8(filename))) LIKE '%.v'
```

Uses the *derived dataset* popular\_content\_name from the SWH graph

**Data**

Data source: **AwsDataCatalog**

Database: **derived\_20220425**

Tables and views: **Create**

Filter tables and views

- Tables (3)
  - path\_counts\_forward\_orisnpretrevdir
  - popular\_content
  - popular\_content\_name
- Views (0)

```
1 select swhid, lower(trim(from_utf8(filename))) from popular_content_name where lower(trim(from_utf8(filename))) LIKE '%.v'
```

SQL Ln 1, Col 1

**Run again** Explain Cancel Clear Create

Reuse query results up to 60 hours ago

Query results Query stats

**Completed** Time in queue: 105 ms Run time: 25.112 sec Data scanned: 329.37 GB

**Results (3,230,368)**

Copy Download results

Search rows

#	swhid	_col1
1	swh:1:cnt:d0b42b2f6049bb3ad5375de8da2ca4470a88143d	design_1_rst_ps7_0_100m_0_sim_netlist.v
2	swh:1:cnt:da86b10e27545466132886ca6c8c2ba2ec727dd3	design_1_auto_pc_0_sim_netlist.v
3	swh:1:cnt:8dedf77644ee8b951e353b0f7541ee0ee0dba557	design_1_auto_pc_0_stub.v
4	swh:1:cnt:b6a517709fd6ca3fdc356251354eb4ae436b1c1e	design_1_processing_system7_0_0_sim_netlist.v
5	swh:1:cnt:d8278ae00e1249f90e653e6eace563d5888b15ca	uart_system_processing_system7_0_0_stub.v
6	swh:1:cnt:06f3b029b61a23ca747dee456967e9f02634cb60	uart_system_processing_system7_0_0_sim_netlist.v
7	swh:1:cnt:cf61ba140738f473ddaf18b81283af0b160fe6f0	const.v
8	swh:1:cnt:4e4605d7064ee7eccde366b6694894b89a671644	constant_folding.v

# Looking for founding principles at Software Heritage

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.

# Findings from [BigCode: The Stack v2](#) and [StarCoder2](#)


Dataset



## The Stack v2

BigCode × Software Heritage  
THE GREAT LIBRARY OF SOURCE CODE

Model



## StarCoder2

built by BigCode

supported by

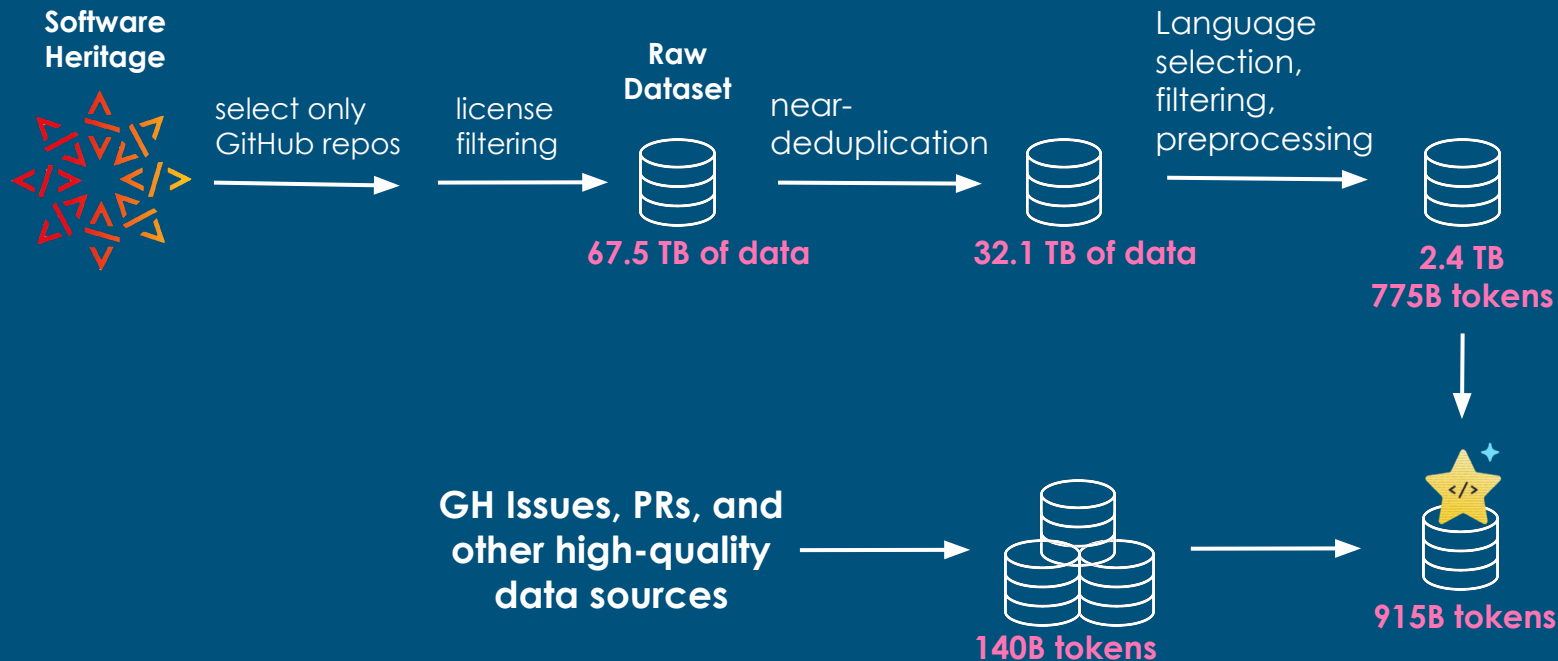
servicenow × 🤗 × NVIDIA

*Released February 28th 2024*

**Yes one can build [the best open LLM for code available](#) while fully adhering to the Software Heritage principles for responsible LLMs, ...  
*and even more: the full training pipeline is made public too!***

# The Stack v2

Data collection pipeline fully open and transparent built by BigCode



# I found my (L)GPL code in your dataset!



Tim Davis  
@DocSparse

@github copilot, with "public code" blocked, emits large chunks of my copyrighted code, with no attribution, no LGPL license. For example, the simple prompt "sparse matrix transpose, cs\_" produces my cs\_transpose in CSparse. My code on left, github on right. Not OK.

```
Terminal
Include "cs.h"
C = A * M;
*cs_transpose (const cs *A, cst values)

csi p, q, j, *Cd, *Cl, n, m, *Ap, *Al, *w;
double *Cx, *Ax;
CS *C;
if (CS_CSC (A)) return (NULL); /* check inputs */
M = A->n; n = A->n; Ap = A->p; Al = A->l; Ax = A->x;
C = cs_spalloc (n, m, Ap [n], values 88 Ax, 0); /* allocate result */
w = cs_calloc (0, sizeof (cst)); /* get workspace */
if (C [1] [w]) return (cs_done (C, w, NULL, 0)); /* out of memory */
Cp = C->p; Cl = C->l; Cx = C->x;
for (p = 0; p < Ap [0]; p++) w [Ap [p]]++; /* row counts */
cs_consum (Cp, M, n); /* row pointers */
for (j = 0; j < n; j++)
for (p = Ap [j]; p < Ap [j+1]; p++)
{
    Cl [q = w [Al [p]]++] = j; /* place A(i,j) as entry C(j,i) */
    if (Cx) Cx [q] = Ax [p];
}
return (cs_done (C, w, NULL, 1)); /* success; free w and return C */
```

SIAM NEWS DECEMBER 2022

Science Policy | December 01, 2022 Print

## Ethical Concerns of Code Generation Through Artificial Intelligence

By Tim Davis and Siva Rajamanickam

Machine learning models that are trained on large corpuses of text, images, and source code are becoming increasingly common. Such models—which are either freely available or accessible for a fee—can then generate their own text, images, and source code. The unprecedented pace of development and adoption of these tools is quite different from the traditional mathematical software development life cycle. In addition, developers are creating large language models (LLMs) for text summarization as well as caption and prompt generation. LLMs are fine-tuned on source code, such as in [OpenAI Codex](#), which yields models that can interactively generate code with minimal prompting. For example, a prompt like "sort an array" produces code one line at a time that a programmer can then either choose to accept or use to generate a match for an entire sort routine.

<https://sinews.siam.org/Details-Page/ethical-concerns-of-code-generation-through-artificial-intelligence>

# The BigCode approach: data inspection and opt out

The screenshot shows a GitHub issue page for 'bigcode-project / opt-out-v2'. The 'Issues' tab is highlighted with a red box. The issue title is 'Opt-out request for nuprl #54' and it is marked as 'Closed'. The issue was opened by 'arjungguha' on Nov 7, 2023. There are three comments:

- arjungguha** 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."
- arjungguha** commented on Nov 7, 2023: "This is a benchmark, and was used in the StarCoder paper. So best not to train on it. :)"
- 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]"

## Members

Users > swayam

```
1 def is_  
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
```

Beware of false positives: *not everything is copyrightable*, e.g. boilerplate, or purely functional code like this one!

Highlighted code was found in the stack.

Source: HF Code Autocomplete (Extension)

Go to stack search

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



# Lessons learned

## 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.

**Transparency is easy:** [SWHID](#) (undergoing ISO standardisation) and Software Heritage  
N.B. : may be mandated by regulations!

**Opt out is complex:** who is *the real right owner*?  
(similar issues to license compliance)



- **Building the training set is complex:**  
e.g. includes **license compliance**  
alike work **at massive scale**
- **Generating attribution information**  
on model output is **more complex**  
than license compliance

**We need a coordinated effort** to ensure fully open models will succeed!

# GENERATIVE AI FOR CODE : THE OPEN ISSUES

Gousios et al. GHTorrent : [GitHub's data from a firehose](#). MSR 2012

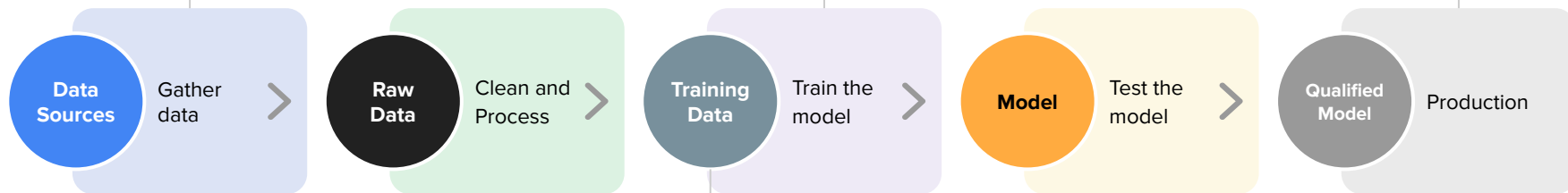
Collect source code, issues, PR, discussions, etc. **is very expensive**. Redoing it over and over again is an **anti-ecological waste**.

Lefevre 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

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

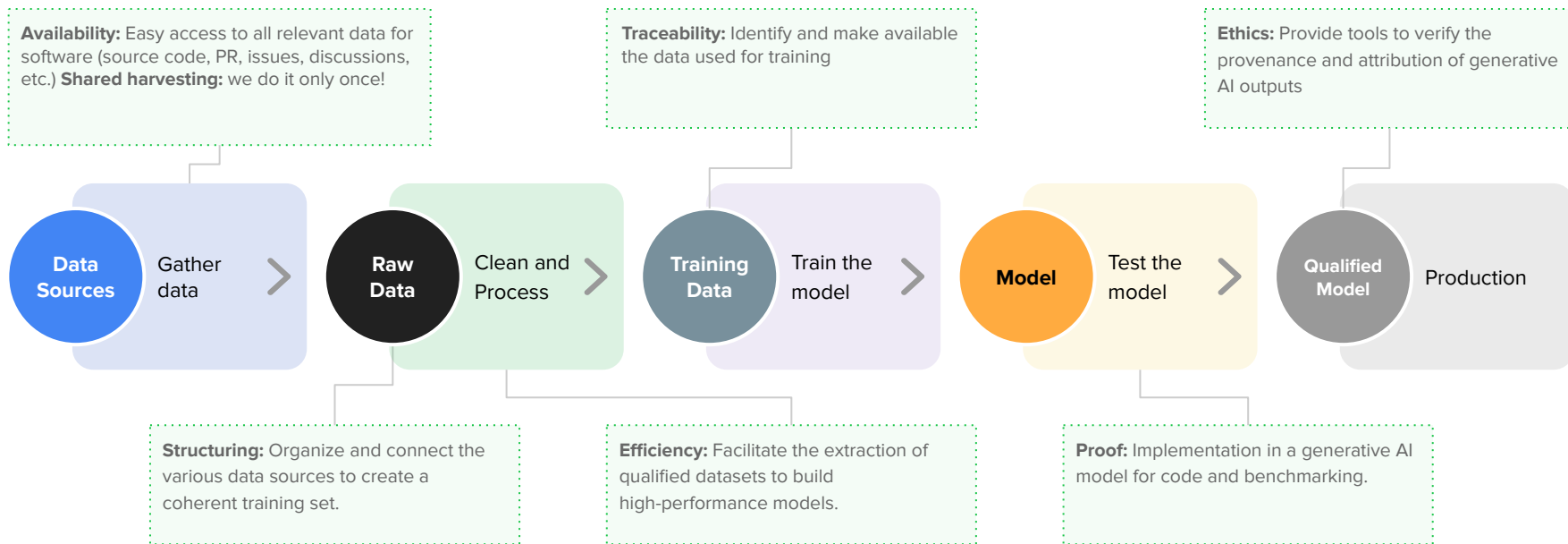
Extracting **quality subsets** should allow to **specialize LLMs** to perform **quality programming and software engineering tasks**.

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

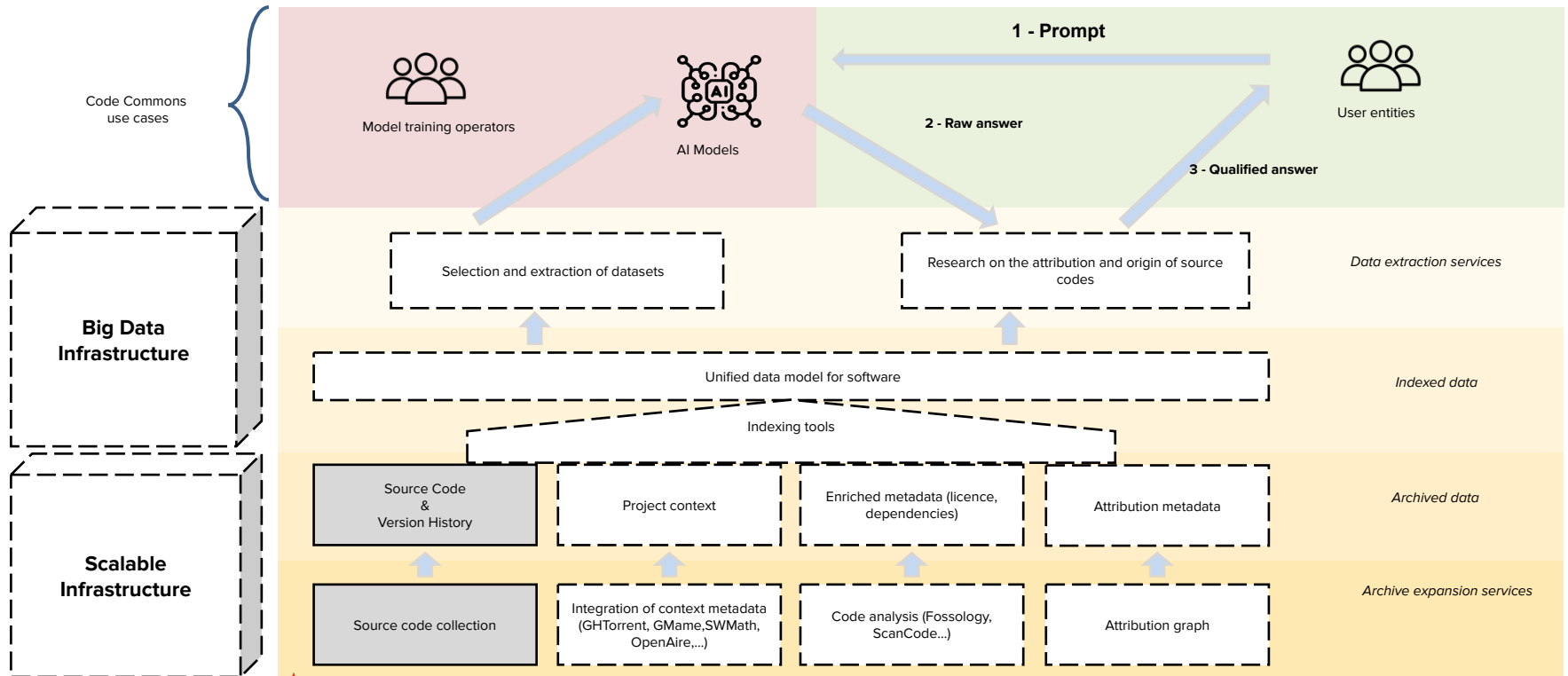
Ledivrec et al. [HyperDiff: Computing Source Code Diffs at Scale](#) ASE 2023

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


# A STEP FORWARD: CODE COMMONS



# CODE COMMONS: bird's eye view

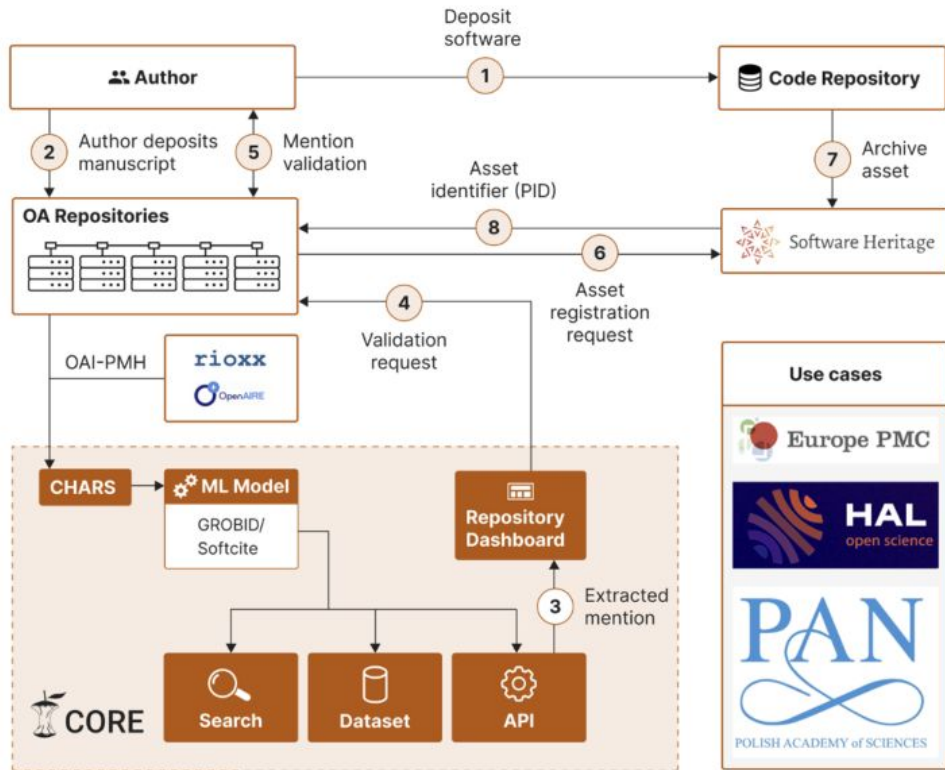


# CODE COMMONS : MEET THE TEAMS

Team	Entity / Person	Expertise
<b>Funded Partners</b>		
 Software Heritage		Universal Software Source Code Archive
 <b>DiverSE</b> <small>RESEARCH IN SOFTWARE VARIABILITY MANAGEMENT</small>		Software engineering, code, programming, languages, software variability management Large-scale software evolution, generative AI for software development
 ALMAnaCH		Automatic linguistic modeling and analysis, and computational humanities
 CEDAR		Analysis and processing of large-scale complex data
DIASI		Natural Language Processing (NLP) Generative AI
DILS		Engineering, Software, and Systems
Software Innovation Lab	 <small>by Modus Create</small>	Machine Learning, Modeling, Natural Language Processing (NLP) Distributed Computing
<b>Subcontracting (budget &lt; 200k€)</b>		
	Philippe Ombredanne	La référence mondiale pour la détection des licences
<b>External contributors</b>		
Emérite Inria	Patrick Valduriez	Cutting-edge expertise in big data management
 <b>UNIVERSITÀ DI PISA</b>	Paolo Ferragina Marco Danelutto	Data compression and text algorithms (ACM Paris Kanellakis Award 2022) Expertise in massively parallel programming HPC
 <b>ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA</b>	Maurizio Gabbriellini	Expertise in machine learning and text similarity
 <b>UNIVERSITÀ DEGLI STUDI DI TORINO</b>	Marco Aldinucci	EuroHPC and expertise in efficient low-level distributed structures

# Related projects

## SoFAIR



# SWH-Sec

## Clear synergies

- HPC Infrastructure
- Project/code metadata

## LLM4Code

### “Défi Inria”

- Reliable and productive code assistants based on LLMs
- 10 Inria teams
- Research project