OPEN SOURCE AI DEFINITION

Online public townhall

May 31, 2024

last updated: May 28, 2024 (MJ)

Community agreements

- **One Mic, One Speaker** -- Please allow one person to speak at a time.
- **Take Space, Make Space** -- If you tend to talk more, we invite you to make space for others to share. If you tend not to share, we invite you to speak up.
- **Kindness** -- This work is hard, but we don't have to be. Gentleness and curiosity help. Those who use insults or hate speech will need to leave the meeting.
- **Forward Motion** -- We advance by focusing on what is possible in the moment and doing it. Obstacles are marked for later discussion, not used to stop the process. If we hit a boulder, we note it on the map and keep walking. We'll come back and unearth it later on.
- **Solution-Seeking** -- This work is so complex that focusing on what won't work will stop it. Suggesting new ideas, options, and proposals is vulnerable, but crucial. All of us are needed to make this work.
- **Anything else?**

OSI's objective for 2024

Open Source AI Definition

Open Source AI Definition
Where Are We Now?

Open Source AI Definition

v.0.0.8

Preamble

Preamble

Why we need Open Source Artificial Intelligence (AI)

Open Source has demonstrated that massive benefits accrue to everyone when we remove the barriers to learning, using, sharing and improving software systems. These benefits are the result of using licenses that adhere to the Open Source Definition. The benefits can be summarized as autonomy, transparency, frictionless reuse, and collaborative improvement.

Everyone needs these benefits in AI. We need essential freedoms to enable users to build and deploy AI systems that are reliable and transparent.

What is Open Source AI

4 Freedoms

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspect its components.
- **Modify** the system for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.

Precondition to exercise these freedoms is to have access to the preferred form to make modifications to the system.

Preferred form to make modifications to machine-learning systems

The preferred form of making modifications for a machine-learning Open Source AI must include:

- **Data information**: Sufficiently detailed information about the data used to train the system, so that a skilled person can recreate a substantially equivalent system using the same or similar data.
  - For example, if used, this would include the training methodologies and techniques, the training data sets used, information about the provenance of those data sets, their scope and characteristics, how the data was obtained and selected, the labeling procedures and data cleaning methodologies.
- **Code**: The source code used to train and run the system.
  - For example, if used, this would include code used for pre-processing data, code used for training, validation and testing, supporting libraries like tokenizers and hyperparameters search code, inference code, and model architecture.
- **Model**: The model parameters.
  - For example, this might include checkpoints from key intermediate stages of training as well as the final optimizer state.

Legal Checklist

Checklist to evaluate machine learning systems

This checklist is based on the paper The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI published Mar 21, 2024.

Table of default required components

Required components	Legal frameworks
Data information	
– Training methodologies and techniques	Available under OSD-compliant license
– Training data scope and characteristics	Available under OSD-compliant license
– Training data provenance (including how data was obtained and selected)	Available under OSD-compliant license
– Training data labeling procedures, if used	Available under OSD-compliant license
– Training data cleaning methodology	Available under OSD-compliant license
Code	
– Data pre-processing	Available under O SI-approved license
– Training, validation and testing	Available under O SI-approved license
– Inference	Available under O SI-approved license
– Supporting libraries and tools	Available under O SI-approved license
Model	
– Model architecture	Available under OSI-approved license
– Model parameters	Available under OSD-conformant terms

The following components are not required as the preferred form of making modifications, but their inclusion in releases is appreciated.

Optional components	Legal frameworks
Data information All data sets, including:	
– Training data sets	Available under OSD-compliant license
– Testing data sets	Available under OSD-compliant license
– Validation data sets	Available under OSD-compliant license
– Benchmarking data sets	Available under OSD-compliant license
– Data card	Available under OSD-compliant license
– Evaluation data	Available under OSD-compliant license

Open Source AI Definition

v.0.0.8

Preamble

4 Freedoms

Legal Checklist
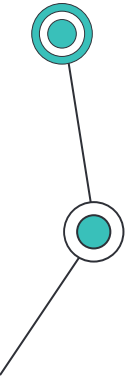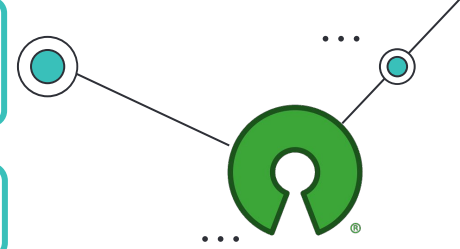
Preamble

Why we need Open Source Artificial Intelligence (AI)

Autonomy, transparency and massive benefits accrue to everyone when you remove the barriers to learning, using, sharing and improving software systems. These benefits are the result of using licenses that adhere to the Open Source Definition. The benefits can be summarized as autonomy, transparency, frictionless reuse, and collaborative improvement.

Everyone needs these benefits in AI. We need essential freedoms to enable users to build and deploy AI systems that are reliable and transparent.

What is Open Source AI

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspect its components.
- **Modify** the system for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.

Precondition to exercise these freedoms is to have access to the preferred form to make modifications to the system.

Preferred form to make modifications to machine-learning systems

The preferred form of making modifications for a machine-learning Open Source AI must include:

- **Data information**: Sufficiently detailed information about the data used to train the system, so that a skilled person can recreate a substantially equivalent system using the same or similar data.
  - For example, if used, this would include the training methodologies and techniques, the training data sets used, information about the provenance of those data sets, their scope and characteristics, how the data was obtained and selected, the labeling procedures and data cleaning methodologies.
- **Code**: The source code used to train and run the system.
  - For example, if used, this would include code used for pre-processing data, code used for training, validation and testing, supporting libraries like tokenizers and hyperparameters search code, inference code, and model architecture.
- **Model**: The model parameters.
  - For example, this might include checkpoints from key intermediate stages of training as well as the final optimizer state.

Checklist to evaluate machine learning systems

This checklist is based on the paper The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI published Mar 21, 2024.

Table of default required components

Required components	Legal frameworks
Data information	
– Training methodologies and techniques	Available under OSD-compliant license
– Training data scope and characteristics	Available under OSD-compliant license
– Training data provenance (including how data was obtained and selected)	Available under OSD-compliant license
– Training data labeling procedures, if used	Available under OSD-compliant license
– Training data cleaning methodology	Available under OSD-compliant license
Code	
– Data pre-processing	Available under OSI-approved license
– Training, validation and testing	Available under OSI-approved license
– Inference	Available under OSI-approved license
– Supporting libraries and tools	Available under OSI-approved license
Model	
– Model architecture	Available under OSI-approved license
– Model parameters	Available under OSD-conformant terms

The following components are not required as the preferred form of making modifications, but their inclusion in releases is appreciated.

Optional components	Legal frameworks
Data information All data sets, including:	
– Training data sets	Available under OSD-compliant license
– Testing data sets	Available under OSD-compliant license
– Validation data sets	Available under OSD-compliant license
– Benchmarking data sets	Available under OSD-compliant license
– Data card	Available under OSD-compliant license
– Evaluation data	Available under OSD-compliant license

Working on review processes for determining if an AI system meets the definition requirements

Open Source AI Definition
How Did We Get Here?
May 2024

Validation Reviewers

We're interested in reviewing about 10 AI systems self-described as open as part of this definition validation phase. Those marked (*) have were reviewed in previous phases.

1. Arctic
1. **Jesús M. Gonzalez-Barahona** Universidad Rey Juan Carlos

2. BLOOM*
2. **Danish Contractor** BLOOM Model Gov. Work Group
3. **Jaan Li** University of Tartu, One Fact Foundation

3. Falcon
1. **Casey Valk** Nutanix
2. **Jean-Pierre Lorre** LINAGORA, OpenLLM-France

4. Grok
1. **Victor Lu** independent database consultant
2. **Karsten Wade** Open Community Architects

5. Llama 2*
1. **Davide Testuggine** Meta
2. **Jonathan Torres** Meta
3. **Stefano Zacchiroli** Polytechnic Institute of Paris
4. **Victor Lu** independent database consultant

6. Mistral
5. **Mark Collier** OpenInfra Foundation
6. **Jean-Pierre Lorre** LINAGORA, OpenLLM-France
7. **Cailean Osborne** University of Oxford, Linux Foundation

7. OLMo
1. **Amanda Casari** Google
2. **Abdoulaye Diack** Google

8. OpenCV*
1. **Rasim Sen** Oasis Software Technology Ltd.

9. Phi-2
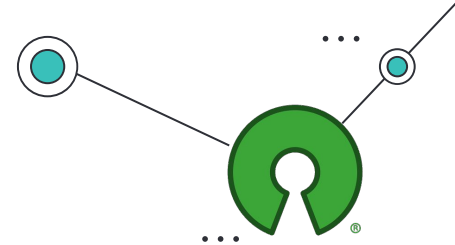3. **Seo-Young Isabelle Hwang** Samsung

10. Pythia*
1. **Seo-Young Isabelle Hwang** Samsung
2. **Stella Biderman** EleutherAI
3. **Hailey Schoelkopf** EleutherAI
4. **Aviya Skowron** EleutherAI

11. T5
5. **Jaan Li** University of Tartu, One Fact Foundation

Open Source AI Definition
Validation phase
v.0.0.8

OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

100% $ % .0 .00 123 Arial 11 B I S A

D12

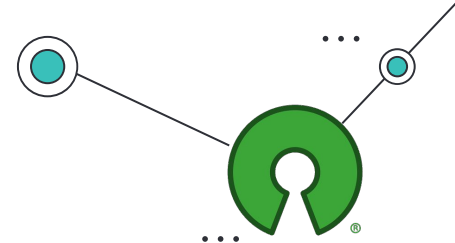|	Component	Legal Framework	C	Legal Document	E	Use for any purpose and without having to ask for permission	Study how the system works and inspect its components	Modification for any purpose, including to change its output	Sharin with or for any
1	**Component**	**Legal Framework**		**Legal Document**		**Arctic Document Analysis**			Freedoms
2	Component definitions: Model Openness Framework	For each component (source: OSAID v. 0.0.8		Paste link to each component's legal document below					
3	Version reviewed: snowflake-arctic-instruct and snowflake-arctic-base (both seems to have the same characteristics) https://huggingface.co/Snowflake/snowflake-arctic-instruct https://huggingface.co/Snowflake/snowflake-arctic-base								
18	Model parameters	Available under OSD-conformant terms		https://huggingface.co/Snowflake/snowfl		Allowed	Allowed	Allowed	Allo
19	**Optional**								
20	**Data Information** All data sets, including:								
21	Training data sets	Available under OSD-compliant license							
22	Testing data sets	Available under OSD-compliant license							
23	Validation data sets	Available under OSD-compliant license							
24	Benchmarking data sets	Available under OSD-compliant license							
25	Data card	Available under OSD-compliant license							
26	Evaluation data	Available under OSD-compliant license				Allowed	Allowed	Allowed	Allo
27	Evaluation results	Available under OSD-compliant license							
28	Other data documentation	Available under OSD-compliant license							
29	**Code**								
30	Code used to perform inference for benchmark tes	Available under OSI-approved license		https://github.com/Snowflake-Labs/snowflake					
31	Evaluation code	Available under OSI-approved license							
32	**Model** All model elements, including:								
33	Model card	Available under OSD-compliant license		https://huggingface.co/Snowflake/snowflake-		Allowed	Allowed	Allowed	Allo
34	Sample model outputs	Available under OSD-compliant license							

+ ☰ 3 Definition Validation -- OLMo ▾ Definition Validation -- Phi-2 ▾ 4 Definition Validation -- T5 ▾ Definition Validation -- TBD ▾ Form F

Open Source AI Definition
Validation phase
v.0.0.8

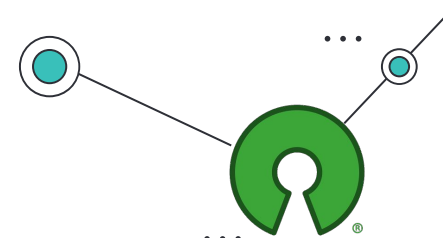It was hard for volunteer reviewers to find required documents to do the review.

OSI: AI Systems Review Workgroups

File   Edit   View   Insert   Format   Data   Tools   Extensions   Help

D12

	Component	Legal Framework		Legal Docume		Arctic Document Analysis		Freedoms
1								
2				Paste link to each compone...gal document below		**Use** for any purpose and without having to ask for permission	**Study** how the system works and inspect its components	**Modification** for any purpose, including to change its output	**Sharin** with o... for any
3	Component definitions: Model Openness Framework	For each component (source: OSAID v. 0.0.8						
	Version reviewed: snowflake-arctic-instruct and snowflake-arctic-base (both seems to have the same characteristics) https://huggi...ace.co/Snowflake/snowflake-arctic-instruct https://huggingface.co/Snowflake/snowflake-arctic-base							
18	Model parameters	Available under OSD-conformant terms		https://huggingface.co...flake/snowfl...	Allowed	Allowed	Allowed	Allow
19	**Optional**							
20	**Data Information** All data sets, including:							
21	Training data sets	Available under OSD-compliant license						
22	Testing data sets	Available under OSD-compliant license						
23	Validation data sets	Available under OSD-compliant license						
24	Benchmarking data sets	Available under OSD-compliant license						
25	Data card	Available under OSD-compliant license						
26	Evaluation data	Available under OSD-compliant license			Allowed	Allowed	Allowed	Allow
27	Evaluation results	Available under OSD-compliant license						
28	Other data documentation	Available under OSD-compliant license						
29	**Code**							
30	Code used to perform inference for benchmark tes	Available under OSI-approved license		https://github.com/Snowflake-Labs/snowflake				
31	Evaluation code	Available under OSI-approved license						
32	**Model** All model elements, including:							
33	Model card	Available under OSD-compliant license		https://huggingface.co/Snowflake/snowflake-...	Allowed	Allowed	Allowed	Allow
34	Sample model outputs	Available under OSD-compliant license						

3 Definition Validation -- OLMo   Definition Validation -- Phi-2   4 Definition Validation -- T5   Definition Validation -- TBD   Form F

Open Source AI Definition
Validation phase
v.0.0.8

This meant a lot of the review analysis was left incomplete

OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

D12

	A	B	C	D	E		G	H	
1	**Component**	**Legal Framework**		**Legal Document**		**Arctic Document Analysis**			Freedoms
2				Paste link to each component's legal document below	**Use** for and with ask for p	urpose aving to sion	**Study** how the system works and inspect its components	**Modification** for any purpose, including to change its output	**Shari** with o for any
3	Component definitions: Model Openness Framework	For each component (source: OSAID v. 0.0.8							
	Version reviewed: snowflake-arctic-instruct and snowflake-arctic-base (both seems to have the same characteristics) https://huggingface.co/Snowflake/snowfla				rctic-instruct https://huggingface.co/Snowflake/snowflake-arctic-base				
18	Model parameters	Available under OSD-conformant terms		https://huggingface.co/Snowfl	Allowed		Allowed	Allowed	Allo
19	**Optional**								
20	**Data Information** All data sets, including:								
21	Training data sets	Available under OSD-compliant license							
22	Testing data sets	Available under OSD-compliant license							
23	Validation data sets	Available under OSD-compliant license							
24	Benchmarking data sets	Available under OSD-compliant license							
25	Data card	Available under OSD-compliant license							
26	Evaluation data	Available under OSD-compliant license			Allowed		Allowed	Allowed	Allo
27	Evaluation results	Available under OSD-compliant license							
28	Other data documentation	Available under OSD-compliant license							
29	**Code**								
30	Code used to perform inference for benchmark tes	Available under OSI-approved license		https://github.com/Snowflake-Labs/snowflake					
31	Evaluation code	Available under OSI-approved license							
32	**Model** All model elements, including:								
33	Model card	Available under OSD-compliant license		https://huggingface.co/Snowflake/snowflake-	Allowed		Allowed	Allowed	Allo
34	Sample model outputs	Available under OSD-compliant license							

3 Definition Validation -- OLMo   Definition Validation -- Phi-2   4 Definition Validation -- T5   Definition Validation -- TBD   Form

Open Source AI Definition
What's Next?
June - October 2024

- Complete validation phase (June 10)
- Resolve comments, release v. 0.0.9 after validation
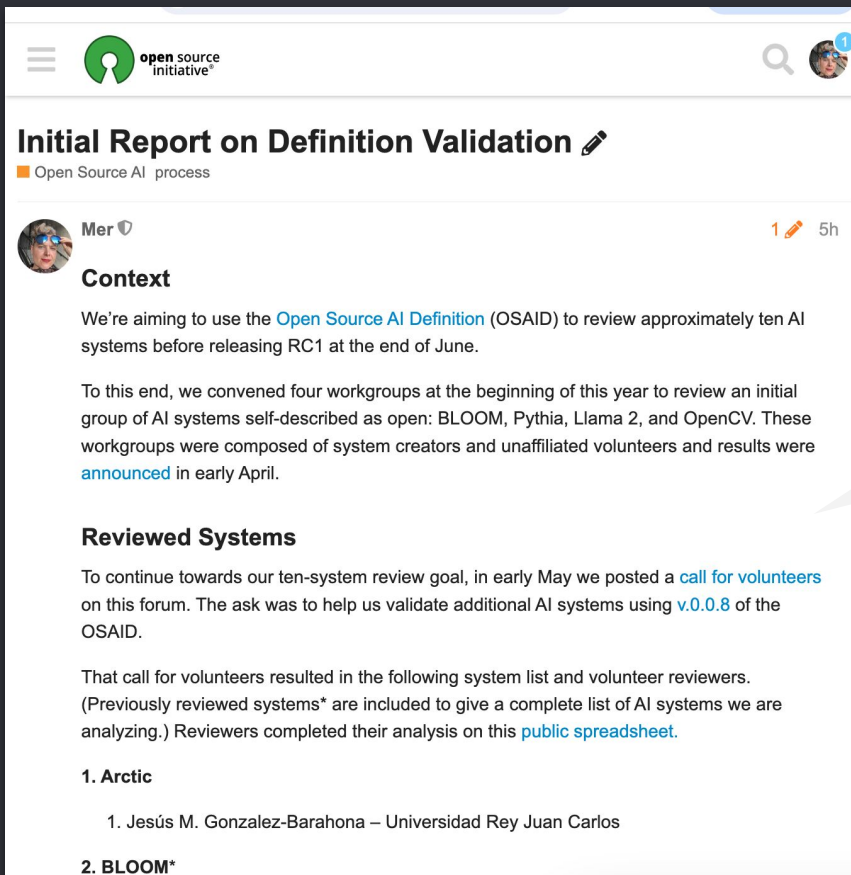- Cut the release candidate with sufficient endorsement

Complete the Validation Phase
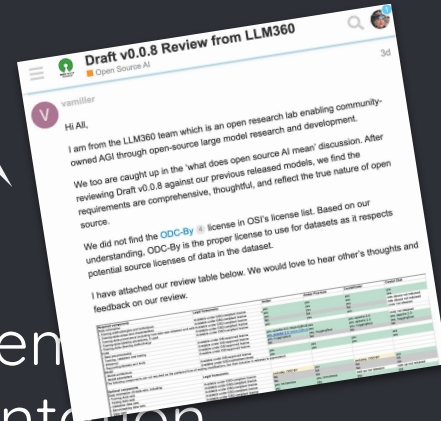
1. Reach out to **AI system creators** to fill in the blanks on their own system by pointing us to correct documentation

2. Invite **volunteers** to also help us fill in these blanks (forum post forthcoming)

3. We're also currently engaging in email and phone debriefs with **reviewers** to better understand the blockers they faced.

Seek Public Feedback on Initial Results

Simplify the Validation Process

2024 Timeline

February	March	April	May	June ...	... October
Call For Volunteers + Activity Feedback and Revision	**Virtual System Review Meetings Begin**	**Virtual System Review Meetings Continue**	**Virtual System Review Meetings END**	**Feedback Informs Content of OSI In-Person Stakeholder Meeting**	**Monthly Virtual Meetings**
Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Townhalls + **PyCon Workshop** (≈ May 17th, Pittsburgh)	Townhall + **Virtual Launch Event** (date TBD)	**Release stable version**
Draft 0.0.5	Draft 0.0.6	Draft 0.0.7 and 8	Draft 0.0.9	RC1	Stable Version

In-Person Meetings

Region	Country	City	Conference	Date
North America	United States	Pittsburgh	✓ **PyCon US**	May 17
Europe	France	Paris	OW2	June 11-12
Africa	Virtual	Virtual	**Sustain Africa**	June
North America	United States	New York	**OSPOs for Good**	July 9 – 11
Asia Pacific	China	Hong Kong	**AI_dev**	August 23
Latin America	Argentina	Buenos Aires	**Nerdearla**	September
Europe	France	Paris	(data governance)	October
North America	United States	Raleigh	**All Things Open**	Oct 27 – 29

The renewed discussion on data

1. The **AWS Open Source** team posted a range of concerns with v 0.0.8, foremost on data.
2. Linux Foundation team recommended adding Data card to the required components. Also they argued that Data preprocessing code is unlikely to be shared if the dataset is not shared, too.

Other relevant posts
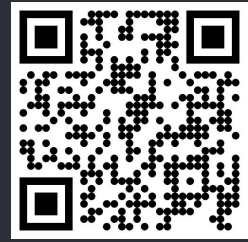
1. The **LLM360** team voluntarily ran their system through the v.0.0.8 <u>review</u> process.

2. Stefano <u>posted</u> on whether and how **OSI** should certify Open Source AI.

Participation Options

- Public **forum**: discuss.opensource.org

- Become an OSI **member**

  - Free or or full

  - SSO with other OSI websites

- Biweekly virtual **townhalls**… like this one!

- **Volunteer** for to fill in the blanks on definition validation (email or DM Mer or Stefano)
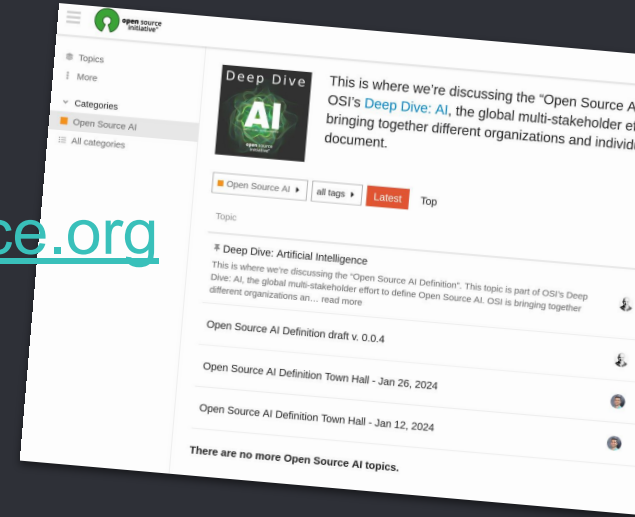
Q & A

Thank you

We realize this is difficult work and we appreciate your help and openness in improving the definitional process.