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 Al Definition

Open Source Al Definition Where Are We Now?

Open Source Al Definition v.0.0.8

Preamble

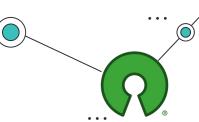
	 Use the system for any purpose and without having to ask for permission.
	 Study how the system works and inspect its components. Modify the system for any purpose, including to change its output.
4 Freedoms	 Modify the system for any purpose, including to change its output. Share the system for others to use with or without modifications, for any purpose.
тпссионы	Precondition to exercise these freedoms is to have access to the preferred form to make
	modifications to the system.
	systems
	The preferred form of making modifications for a machine-learning Open Source Al must include:
	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 same or similar data.
	Por example, if used, this would include the training methodologies and techniques
	the training data sets used, information about the provenance of those data sets 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, cod
	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 tra
	as well as the final optimizer state.
	Checklist to evaluate machine learning syste
	This checklist is based on the paper The Model Openness Framework: Promoting
	Completeness and Openness for Reproducibility, Transparency and Usability in Al published Mar 21, 2024.
	Table of default required components
	Required components Legal frameworks
	Data information
	Training methodologies and techniques Available under OBD- compliant license
	Training data scope and characteristics Available under OSD- compliant license
	Training data provenance (including how data was obtained and selected) 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 0 II-appro license
Logal Chacklist	Training, validation and testing Konse Konse
Legal Checklist	- Inference Available under 0 81-appro license
U	- Supporting libraries and tools Available under 0 SI-appro license
	Model
	Model architecture Model architecture license
	- Model parameters Available under OSD- conformant terms
	The following components are not required as the preferred form of making modification their inclusion in releases is appreciated.
	The following components are not required as the preferred form of making modification their inclusion in releases is appreciated. Optional components Legal frameworks
	their inclusion in releases is appreciated.
	their inclusion in releases is appreciated. Optional components Legal frameworks
	their inclusion in releases is approciated. Optional components Legal frameworks Data Information Af data sets, including: Technologic data sets and the sets and
	Noticitativi in indexes is agreeded. Updato at comparison in the second
	Hari inclusion in inducesis a generative. Section differentiation Air data este, including - Tailong data este, including - Tailong data este - Tailong
	Hairi Indexia Nagaragan Karaka Kar Data Internation At data sets, Indukting Thirting data sets Thirting dat

Preamble

lov Al systems that are re What is Open Source Al

Why we need Open Source Artificial Intelligence (AI)

Open Source has demonstrated that massive benefits accrue to everyone when you remove Open address take settementationed takes in address address of the settement poor tensore the barriers takes address address address address address of the settement poor tensore takes address address address to the Open Source Definition. The benefits can be summarized as automorp, transparency, frictionless reuse, and collaborative improvement. Everyone needs these benefits in Al. We need essential freedoms to enable users to build and



Open Source Al Definition

4 Freedoms	 Use the system for any purpose and althout having Study how the system works and indepect in company Modify its system for any purpose, cluicking to the Starts the system for eithers to as with or arbitrary 	onents. ange its output. modifications, for any purpose.	
	systems		,
	The preferred form of making modifications for a machin	ne-learning Open Source Al must	
	 Visidate Otto MONTRADIO, Torficienty deviated information: points, you have a silicat german can expressing a solution on caracter, and caracter and the solution of the solution the caracter, and caracter and the solution of the solution the caracter, and caracter and the solution of the solution the solution of the solution of the solution of the solution the solution of the solution of the solution of the solution the caracter, and constant caracter and not mit by par- in the solution of the solution of the solution of the solution the solution of the solution of the solution of the solution the solution of the solution of the solution of the solution the solution of the solution of the solution of the solution the solution of the solution of the solution of the solution of the solution the solution of the solution of	stantially equivalent system using the kining methodologies and techniques, he proverance of those data acts, their stander and selected, the labeling rstem. self for pre-processing data, code ting libraries like tokenizers and and model architecture.	
	Checklist to evaluate maching	ne learning systems	
	This checklist is based on the paper The Model Opan Completeness and Openness for Reproducibility, Tran- published Mar 21, 2024.	sess Framework: Promoting sparency and Usability in Al	Working on review
	Table of default required components		
	Required components Data Information	Legal frameworks	processes for
		Available under OSD-	
	- Training methodologies and techniques	compliant license Available under OSD-	determining if an
	Training data scope and characteristics Training data provenance (including how data was	compliant license Available under OSD-	AI system meets
	obtained and selected)	compliant license Available under OSD-	
	- Training data labeling procedures, if used	compliant license	the definition
	- Training data cleaning methodology	Available under OSD- compliant license	
	Code		requirements
	- Data pre-processing	Available under OSI-ap proved license	
	- Training, validation and testing	Available under OSI-ap proved license	
Legal Checklist	- Inference	Available under OSI-approved	
	- Supporting libraries and tools	Available under OSI-approved	
	Model	license	
	- Model architecture	Available under OSI-approved	
		license Available under OSD-	
	Model parameters The following components are not required as the prefer	conformant terms	
	their inclusion in releases is appreciated.		
	Optional components Data information All data sets, including:	Legal frameworks	
	Training data sets	Available under OSD-compliant	
	- Testing data sets	license 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	

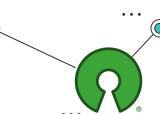
Preamble

Why we need Open Source Artificial Intelligence (AI)

What is Open Source Al

Everyone needs these benefits in Al. We need essential freedoms to enable users to build and

its accrue to everyone when you remove ig software systems. These benefits are the arce Definition. The benefits can be s reuse, and collaborative improvement.



Open Source Al 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

 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*

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

6. Mistral

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

7. OLMo

- 1. **Amanda Casari** Google
- 2. **Abdoulaye Diack** Google

8. OpenCV*

 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** EleutherAl
- 3. Hailey Schoelkopf EleutherAl
- 4. **Aviya Skowron** EleutherAl

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	🛠 🖻 🗠 s Extensions Help				S Share -	*
C	くちさ骨骨 100% - \$ % .0.	00 123 Arial 🝷 - 11 +	B I ÷ <u>A</u> È. ⊞ 53 → Ξ	• <u>↓</u> • ¢ • <u>↓</u> •	☞ ∄ ⊪ Ÿ 📾	- Σ	^
012	✓ ∫ fx						
	A	В	D E	F	G	н	
1	Component	Legal Framework	Legal Document		Arctic Document Anal	ysis Fr	reedom
2	Component definitions: Model Openness Framework	For each component (source: OSAID v. 0.0.8	Paste link to each component's legal	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	Sha with for a
3	Version reviewed: snowflake-arctic-instruct and snowfla				uct https://huggingface.co/S	nowflake/snowflake-arct	tic-base
10	would parameters		nups.//nuggingrace.co/onownake/snowna	Allowed •	Allowed	Allowed	A
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 •	Al
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 -	A
34	Sample model outputs	Available under OSD-compliant license		• • • • •	• Horiou	· · · · ·	

+

Open Source AI Definition Validation phase

v.0.0.8

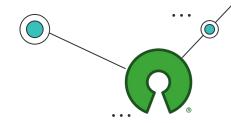
It was hard for volunteer reviewers to find required documents to do the review.



C	२ ५ ट 🛱 🖥 100% ▾ 💲 % .ºੵ .	00 123 Arial ▼ - 11 +	в <u>г ÷ </u> А 🍬 🗄	3	E	<u>• ↓</u> • ₽ • <u>A</u> •	cə (±).		- Σ		^
12	✓ ∫ fx										
	A	в	D		E	F	G		н		
1	Component	Legal Framework	Legal Docume				Arctic Docu	iment Ana	lysis	Fre	edom
2	Component definitions: Model Openness Framework	For each component (source: OSAID v. 0.0.6	Paste link to each compone document below	gal		Use for any purpose and without having to ask for permission	Study how th works and ins components		Modification for purpose, includin change its output	g to	Shar with for a
3	Version reviewed: snowflake-arctic-instruct and snowfla					ke/snowflake-arctic-inst	ruct https://hugg	gingface.co/	Snowflake/snowfla	ke-arcti	c-base
•	would parameters		mups.//nuggingrace.co/c	/IIANC/SI	OWI	Allowed 👻	Allowed	•	Allowed	•	All
9	Optional										
20	Data Information All data sets, including:	Ausilable under OOD somelient lienes									
1	Training data sets	Available under <u>OSD</u> -compliant license						•		•	
2	Testing data sets	Available under <u>OSD</u> -compliant license				•		•		•	
3	Validation data sets	Available under <u>OSD</u> -compliant license Available under <u>OSD</u> -compliant license				•		•		•	
	Benchmarking data sets Data card	Available under OSD-compliant license				•		•		•	
25 26		Available under OSD-compliant license	/			✓	Allowed	•	Allowed	•	
27		Available under OSD-compliant license				Allowed -	Allowed	•	Allowed	•	Alle
.7		Available under OSD-compliant license				•		•		• •	
29	Code							•		•	
0	Code used to perform inference for benchmark tes	Available upder OSLapproved license	https://github.com/Snowflake	Lobe/eno	uflok			•		•	
11	•	Available under OSI-approved license	nups://github.com/Showilake	Labs/snov	VIIAK			•		-	
2	Model All model elements, including:	Available under OSPapproved ilcense				•		•		•	
33	Model card	Available under OSD-compliant license	https://huggingface.co/Snowf	ako/spoud	lako	Allowed -	Allowed	•	Allowed	•	Allo
34	Sample model outputs	Available under OSD-compliant license	https://httpg/ngiace.co/Showi	are/snow	iane-	Allowed	Allowed	-	Allowed	-	Allo

Open Source AI Definition Validation phase

This meant a lot of the review analysis was left incomplete



v.0.0.8

C	Ҳちさ骨骨 100% ▼ \$ % .ºੵ ·	.00 123 Arial ▼ - 11 +	B <i>I</i> ≎ <u>A</u> 	• '₹	•		Ξ * Σ	^
D12	✓ ∫ĵx							
	A	в	D E			G	н	
1	Component	Legal Framework	Legal Document			Arctic Document A	na <mark>l</mark> ysis F	Freedom
2	Component definitions: Model Openness Framework	For each component (source: OSAID v. 0.0.8	Paste link to each component's legal document below	Use for and with ask for pu	urpose aving to sion	Study how the system works and inspect its components	Modification for any purpose, including to change its output	Sha with for a
3	Version reviewed: snowflake-arctic-instruct and snowfla						o/Snowflake/snowflake-arc	
10	wouer parameters	Available under OSD-contormant terms	nups.//nuggingrace.co/onownake/snowna	Allowea	•	Allowed	- Allowed -	• All
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	- All
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	All
34	Sample model outputs	Available under OSD-compliant license			•			•

Open Source Al 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

Thanks, LLM360 team!

Draft v0.0.8 Review from LLM360

yn the LLM360 team which is an open research lab ei

can many use submany reast which is an upon research and developm owned AGI through open-source large model research and developm v0.0.8 against our previous released models, we find the up in the 'what does open source AI mean' dis hensive, thoughful, and reflect the true nature of open

not find the ODC-By \circledast license in OSI's license list. Based on our ing. OOC-By is the proper license to use for datasets as it respect ses of data in the dataset.

. We would love to hear other's

- Reach out to Al system creators to 1. 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



Initial Report on Definition Validation 🖋

Open Source Al process



1 🤌 5h

Please

comment on the initial report of

our validation

process.

We're aiming to use the Open Source AI Definition (OSAID) to review approximately ten AI systems before releasing RC1 at the end of June.

To this end, we convened four workgroups at the beginning of this year to review an initial group of AI systems self-described as open: BLOOM, Pythia, Llama 2, and OpenCV. These workgroups were composed of system creators and unaffiliated volunteers and results were announced in early April.

Reviewed Systems

To continue towards our ten-system review goal, in early May we posted a call for volunteers on this forum. The ask was to help us validate additional AI systems using v.0.0.8 of the OSAID.

That call for volunteers resulted in the following system list and volunteer reviewers. (Previously reviewed systems* are included to give a complete list of AI systems we are analyzing.) Reviewers completed their analysis on this public spreadsheet.

1. Arctic

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

2. BLOOM*

Simplify the Validation Process

SAMPLE OSAID evaluation of File Edit View Insert Forma	card ☆ @ ↔ t Tools Extensions Zotero Help					
< 5 ♂ 母 ♣ 툼 100% ·	• Normal text • Arial • $-$ 11 + B $I \cup A$? : $?$ •	We're exp replacing				
← Outline Evaluation card for Falcon (SAM Analysis of required components Data information Results Evaluator's opinion	Outline Defined as: The source code used to train and run the system. Evaluation card for Falcon (SAM For example, if used, this would include Analysis of required components code used for pre-processing data, code used for training, validation and testing, supporting libraries like tokenizers and hyperparameters search code, inference code, model architecture. 					
 Code: Links to Code Documents Evaluator's opinion Model: Results Evaluator's opinion 	Links to Code Documents https://huggingface.co/tiiuae/falcon-7b#software https://huggingface.co/blog/falcon#inference "https://huggingface.co/tiiuae/falcon-7b/tree/main https://huggingface.co/models?other=custom_code (custom code noted)" Evaluator's opinion Does it include all of the code required (list above)? Yes/No Are all available under an OSD-approved license? Yes/No 					

ploring the neet with ation this €.

2024 Timeline

System testing work stream

Stakeholder consultation work stream

Release schedule

February	March	April	Мау	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	0W2	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	Al_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

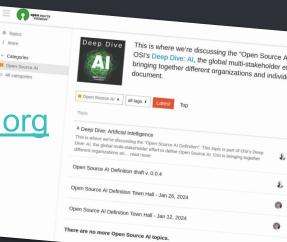
- The AWS Open Source team posted a range of <u>concerns</u> with v 0.0.8, foremost on data.
- 2. Linux Foundation team recommended adding Data card to the required components. Also they argued that Data preprocessing code is unlikely to be shared if the dataset is not shared, too.

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 \bigcirc 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)





Thank you

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