OPEN SOURCE AI DEFINITION

Online public townhall

May 3, 2024

last updated: May 2, 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

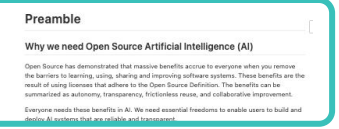
Open Source AI Definition

Open Source AI Definition
Where Are We Now?

Open Source AI Definition **Elements**

v.0.0.8

Preamble



Preamble

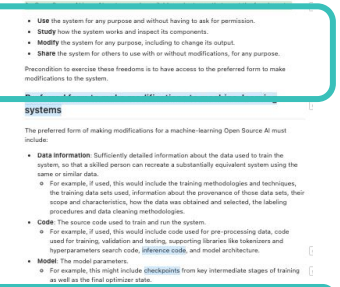
Why we need Open Source Artificial Intelligence (AI)

Open Source has demonstrated that massive benefits accrue to everyone when 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

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.

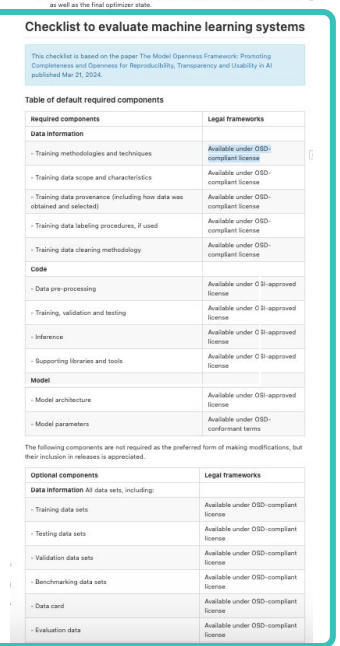
Precondition to exercise these freedoms is to have access to the preferred form to make modifications to the system.

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 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

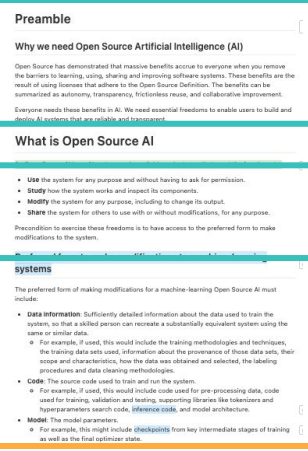
The following components are not required as the preferred form of making modifications, but their inclusion in releases is appreciated.

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

Open Source AI Definition **Elements**

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 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.

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.

4 Freedoms

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 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

Now feature complete for required and optional components

Open Source AI Definition
How Did We Get Here?

Open Source AI Definition

The 4 Freedoms for AI

Fall 2023

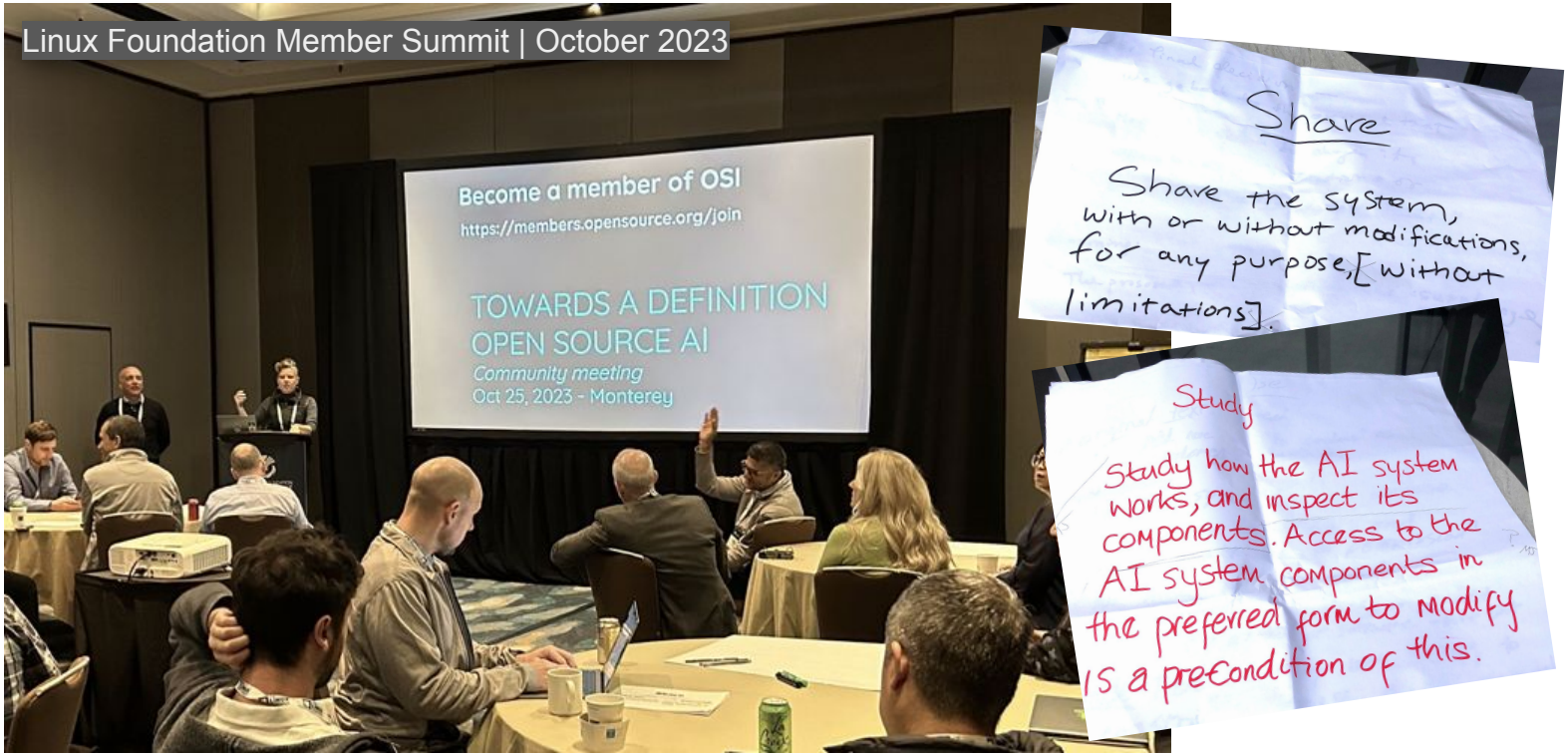
The 4 Freedoms for AI

Use • Study • Modify • Share

What should these open source principles mean for artificial intelligence?

Co-Design Workshop: Raleigh

Co-Design Workshop: Monterey



Co-Design Workshop: Addis Ababa

Open Source AI Definition

The 4 Freedoms for AI

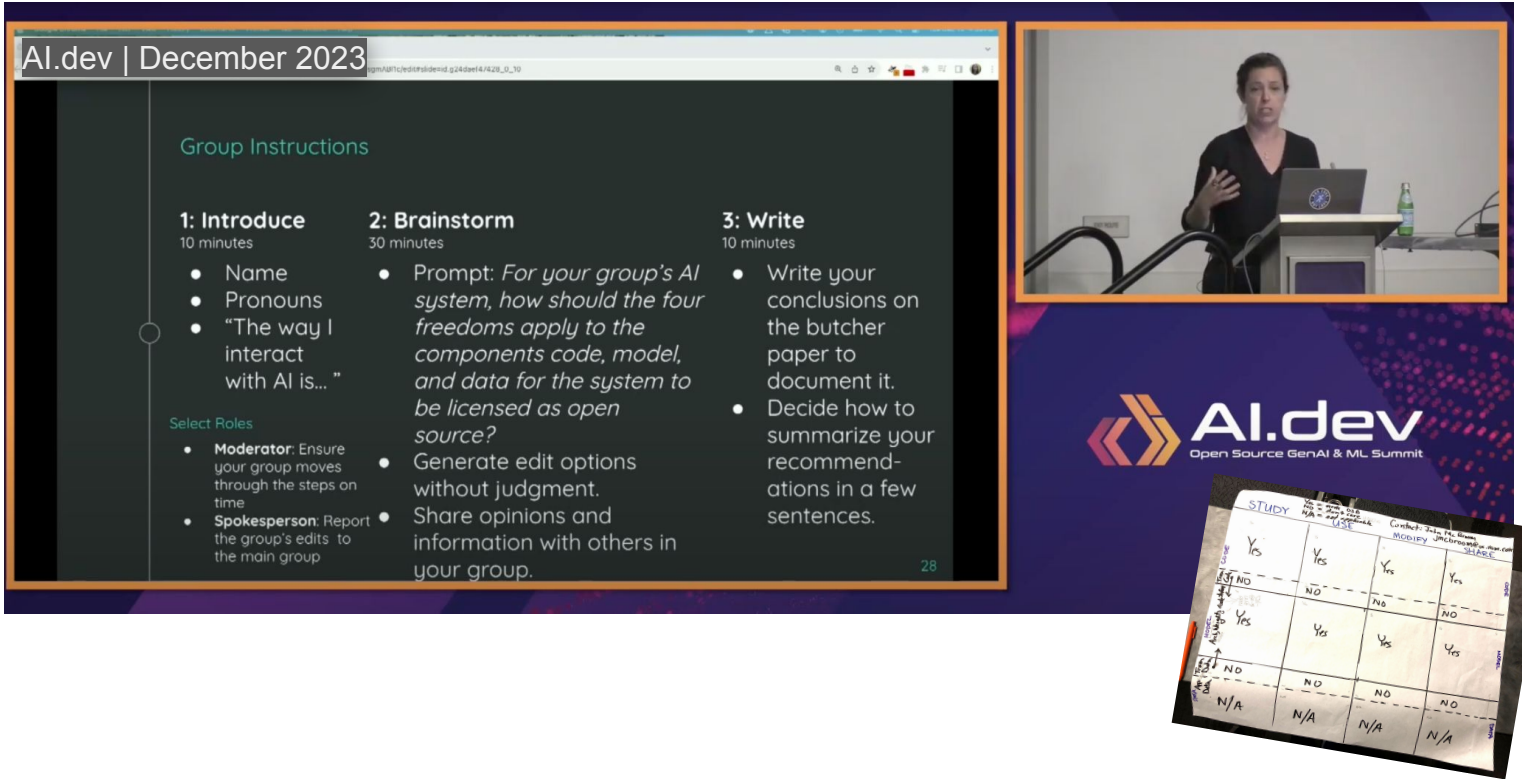
1. **Use** the system for any purpose and without having to ask for permission.
2. **Study** how the system works and inspect its components.
3. **Modify** the system for any purpose, including to change its output.
4. **Share** the system for others to use with or without modifications, for any purpose.

Open Source AI Definition
Required Components
Winter 2023-24

Required Components for Open Source AI

What components must be open in order for an AI system to be used, studied, modified, and shared?

Co-Design Workshop: San Jose



Virtual Workgroups

Selected to represent a diversity of approaches to AI openness:

1. **Llama 2**: commercial project, accompanied by limited amount of science and with a restrictive license
2. **BLOOM**: open science project, with lots of details released but shared with a restrictive license
3. **Pythia**: open science project, with a permissive license
4. **OpenCV**: open source project, with ML components outside of the generative AI space

Workgroup Members

To achieve better global representation, we conducted outreach to Black, Indigenous, and other people of color, particularly women and individuals from the Global South.

Llama 2

1. **Bastien Guerry** DINUM, French public administration
2. **Ezequiel Lanza** Intel
3. **Roman Shaposhnik** Apache Software Foundation
4. **Davide Testuggine** Meta
5. **Jonathan Torres** Meta
6. **Stefano Zacchiroli** Polytechnic Institute of Paris
7. **Mo Zhou** Debian, Johns Hopkins University
8. **Victor Lu** independent database consultant

BLOOM

1. **George C. G. Barbosa** Fundação Oswaldo Cruz
2. **Daniel Brumund** GIZ FAIR Forward - AI for all
3. **Danish Contractor** BLOOM Model Gov. WG
4. **Abdoulaye Diack** Google
5. **Jaan Li** University of Tartu, Phare Health
6. **Jean-Pierre Lorre** LINAGORA, OpenLLM-France
7. **Ofentse Phuti** WiMLDS Gaborone
8. **Caleb Fianku Quao** Kwame Nkrumah University of Science and Technology, Kumasi

Pythia

1. **Seo-Young Isabelle Hwang** Samsung
2. **Cailean Osborne** University of Oxford, Linux Foundation
3. **Stella Biderman** EleutherAI
4. **Justin Colannino** Microsoft
5. **Hailey Schoelkopf** EleutherAI
6. **Aviya Skowron** EleutherAI

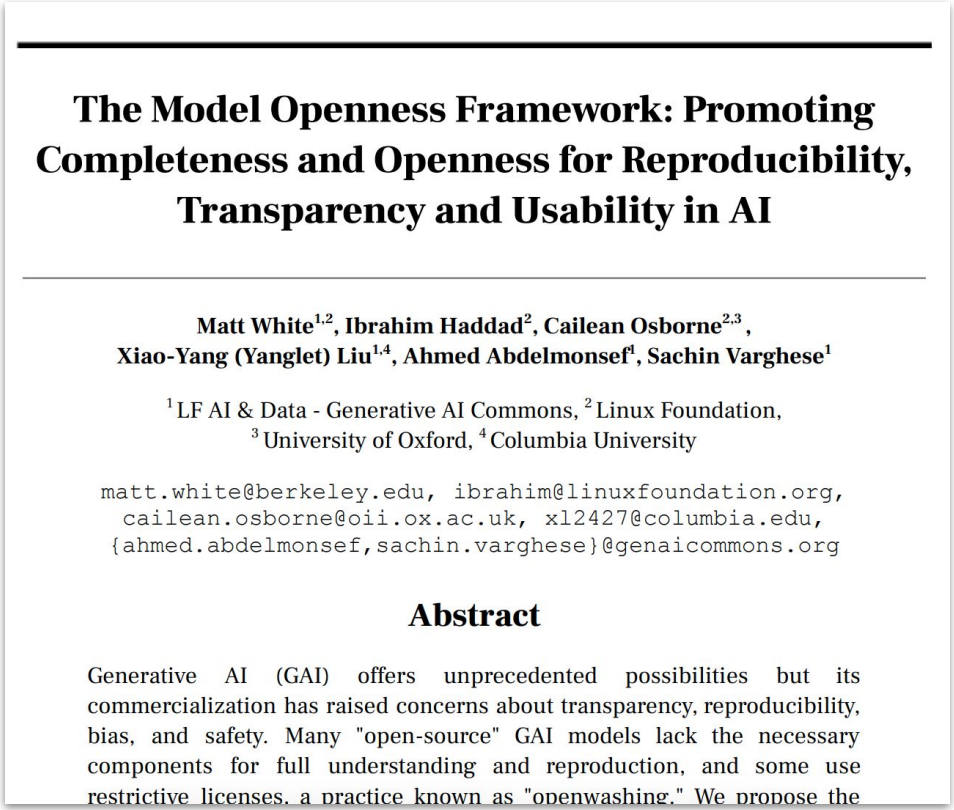
Over 50% of all workgroup participants are people of color.

OpenCV

1. **Rahmat Akintola** Cubeseed Africa
2. **Ignatius Ezeani** Lancaster University
3. **Kevin Harerimana** CMU Africa
4. **Satya Mallick** OpenCV
5. **David Manset** ITU
6. **Phil Nelson** OpenCV
7. **Tlamelo Makati** WiMLDS Gaborone, Technological University Dublin
8. **Minyechil Alehegn Tefera** Mizan Tepi University
9. **Akosua Twumasi** Ghana Health Service
10. **Rasim Sen** Oasis Software Technology Ltd.

Workgroups: Required Component Selection

Component List



The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI

Matt White[1,2], Ibrahim Haddad[2], Cailean Osborne[2,3], Xiao-Yang (Yanglet) Liu[1,4], Ahmed Abdelmonsef[1], Sachin Varghese[1]

[1] LF AI & Data - Generative AI Commons, [2] Linux Foundation, [3] University of Oxford, [4] Columbia University

matt.white@berkeley.edu, ibrahim@linuxfoundation.org, cailean.osborne@oii.ox.ac.uk, xl2427@columbia.edu, {ahmed.abdelmonsef,sachin.varghese}@genaicommons.org

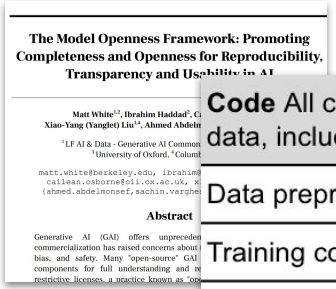
Abstract

Generative AI (GAI) offers unprecedented possibilities but its commercialization has raised concerns about transparency, reproducibility, bias, and safety. Many "open-source" GAI models lack the necessary components for full understanding and reproduction, and some use restrictive licenses, a practice known as "openwashing." We propose the

Available on arXiv

Workgroups: Required Component Selection

Component List


The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI

Abstract

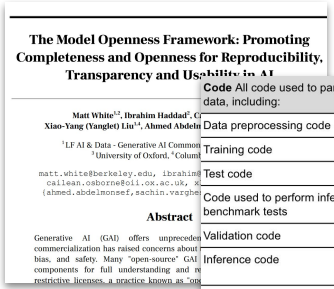
Component Votes

Code All code used to parse and process data, including:	Required to Use?	Required to Study?	Required to Modify?	Required to Share?
Data preprocessing code		SZ	SZ EL	
Training code		SZ	SZ	
Test code				
Code used to perform inference for benchmark tests				
Validation code			SZ	
Inference code	SM EL DT SM JT SZ		SZ	SZ
Evaluation code				
Other libraries or code artifacts that are part of the system, such as tokenizers and hyperparameter search code, if used.	BG,EL, SM, SZ	SZ	SZ	SZ

Example: Llama 2 Workgroup

Workgroups: Required Component Selection

Component List



The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI

Matt White[1,2], Ibrahim Haddad[2], C
Xiao-Yang (Yanglet) Liu[1,4], Ahmed Abdelm

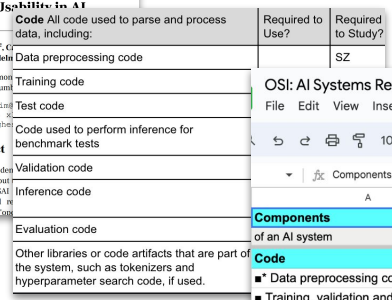
[1]LF AI & Data - Generative AI Common
[3]University of Oxford, [4]Columb

matt.white@berkeley.edu, ibrahim
cailean.osborne@oii.ox.ac.uk,
(ahmed.abdelmonsef,sachin.vargh

Abstract

Generative AI (GAI) offers unpreceden
commercialization has raised concerns about
bias, and safety. Many "open-source" GAI
components for full understanding and re
restrictive licenses, a practice known as "op

Component Votes



Code All code used to parse and process data, including:	Required to Use?	Required to Study?
Data preprocessing code		SZ
Training code		
Test code		
Code used to perform inference for benchmark tests		
Validation code		
Inference code		
Evaluation code		
Other libraries or code artifacts that are part of the system, such as tokenizers and hyperparameter search code, if used.		

Vote Compilation

OSI: AI Systems Review Workgroups

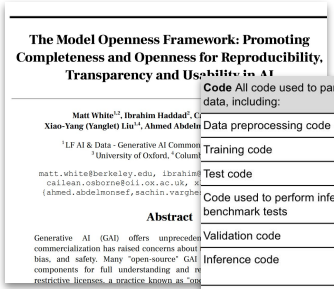
File Edit View Insert Format Data Tools Extensions Help

Components

	A	B	C	D	E	F	G	H	I	J	K	L	M
	Components	**Recommendation**		**Rationale**		**Total**		**Vote Tally** (MOF* update)					**Legend**
	of an AI system	Should it be required?		Why should it be required?		Votes		Study	Use	Modify	Share		Recommmendation key
	Code	last update: 2/26/24 (MJ)											
	■* Data preprocessing code	Lean yes	→	Likely required to study and r	→	29	→	17	-2	13	1		Yes = Required (≥2μ* votes)
	■ Training, validation and testing code	Yes	→	Required to study	→	39	→	24	2	13	0		Lean Yes = Likely required (<2μ-1.5μ votes)
	Test code	[combined into categor	→	Necessary for study, maybe s	→	4	→	4	0	0	0		Maybe = Possibly required (<1.5μ-μ votes)
	Validation code	[combined into categor	→	Necessary for study, maybe r	→	2	→	2	0	1	-1		Lean No = Likely not required (<μ - .5μ votes)
	■ Inference code	Yes	→	Possibly required to use and	→	39	→	11	13	7	8		No = Not required (<.5μ votes)
	■ Evaluation code	Lean no	→	Possibly required to study	→	15	→	10	2	2	1		μ = average votes per component
	Code used to perform inference for bench	No		Likely not required to study		3		6	1	1	-5		As of 2/29/24 μ =
	Data												16
	■ Datasets	Maybe	→	Various datasets possibly req	→	17	→	12	0	5	1		
	▸ Training datasets	Maybe	→	Possibly required for study	→	20	→	13	1	6	0		
	▸ Testing datasets	Maybe	→	Possibly required for study	→	19	→	14	1	4	0		
	▸ Validation datasets	Lean no	→	Likely not required for study	→	13	→	9	-1	5	0		
	▸ Benchmarking datasets	Lean no	→	Likely not required for study	→	15	→	10	-1	4	2		
	■ Data card	No	→	Not required for study	→	1	→	6	-3	-1	-1		
	Evaluation metrics and results	[split into data and resu	→	Not required for study	→	-1	→	4	-3	-1	-1		
	■ Evaluation data	No	→	Not required for study	→	3	→	7	-3	0	-1		
	■ Evaluation results	No	→	Not required for study	→	5	→	9	-3	0	-1		
	All other data documentation	Lean no	→	Possibly required for study	→	13	→	10	-1	4	0		

Workgroups: Required Component Selection

Component List



The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI

Matt White[1,2], Ibrahim Haddad[2], C...
Xiao-Yang (Yanglet) Liu[3,4], Ahmed Abdelm...

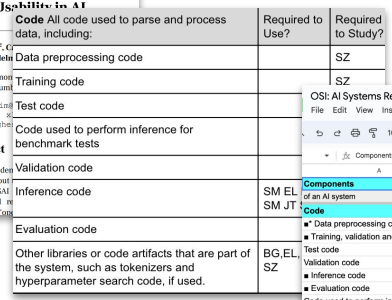
[1]LF AI & Data - Generative AI Common...
[3]University of Oxford, [4]Columb...

matt.white@berkeley.edu, ibrahim...
cailean.osborne@oii.ox.ac.uk, ...
{ahmed.abdelmonsef,sachin.vargh...

Abstract

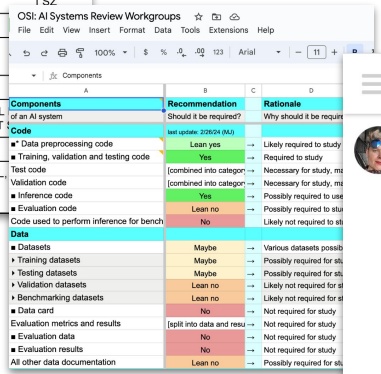
Generative AI (GAI) offers unpreceden...
commercialization has raised concerns about ...
bias, and safety. Many "open-source" GAI ...
components for full understanding and re...
restrictive licenses, a practice known as "ope...

Component Votes



Code All code used to parse and process data, including:	Required to Use?	Required to Study?
Data preprocessing code		SZ
Training code		SZ
Test code		
Code used to perform inference for benchmark tests		
Validation code		
Inference code		SM EL SM JT
Evaluation code		
Other libraries or code artifacts that are part of the system, such as tokenizers and hyperparameter search code, if used.		BG,EL SZ

Vote Compilation

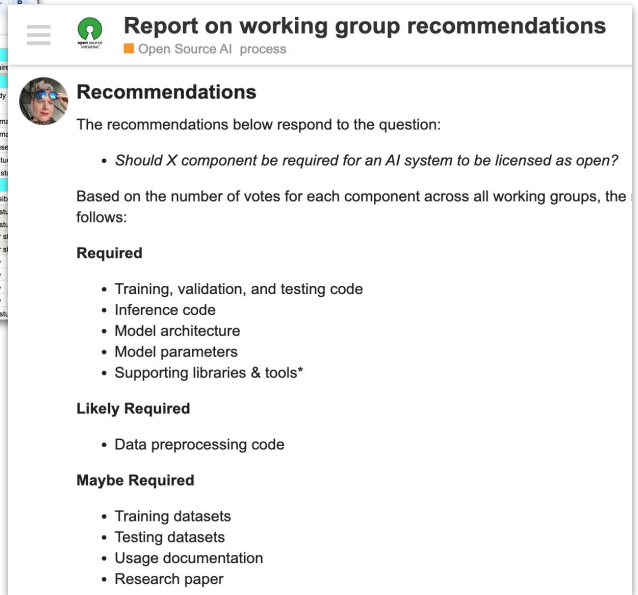


OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

Components	Recommendation	Rationale
of an AI system	Should it be required?	Why should it be require...
Code	last update: 2/26/24 (MJ)	
▪ Data preprocessing code	Lean yes	Likely required to study
▪ Training, validation and testing code	Yes	Required to study
Test code	[combined into categor...	Necessary for study, ma...
Validation code	[combined into categor...	Necessary for study, ma...
▪ Inference code	Yes	Possibly required to use...
▪ Evaluation code	Lean no	Possibly required to stu...
Code used to perform inference for bench...	No	Likely not required to st...
Data		
▪ Datasets	Maybe	Various datasets possib...
▸ Training datasets	Maybe	Possibly required for st...
▸ Testing datasets	Maybe	Possibly required for st...
▸ Validation datasets	Lean no	Likely not required for s...
▸ Benchmarking datasets	Lean no	Likely not required for s...
▪ Data card	No	Not required for study
Evaluation metrics and results	[split into data and resu...	Not required for study
▪ Evaluation data	No	Not required for study
▪ Evaluation results	No	Not required for study
All other data documentation	Lean no	Possibly required for st...

Recommendation Report



Report on working group recommendations

Open Source AI process

Recommendations

The recommendations below respond to the question:

- *Should X component be required for an AI system to be licensed as open?*

Based on the number of votes for each component across all working groups, the follows:

Required

- Training, validation, and testing code
- Inference code
- Model architecture
- Model parameters
- Supporting libraries & tools*

Likely Required

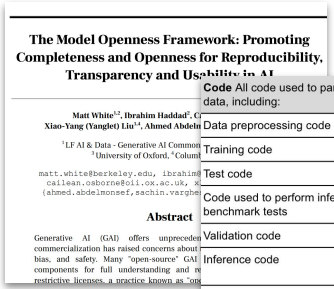
- Data preprocessing code

Maybe Required

- Training datasets
- Testing datasets
- Usage documentation
- Research paper

Workgroups: Required Component Selection

Component List



The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI

Matt White[1,2], Ibrahim Haddad[2], C…
Xiao-Yang (Yanglet) Liu[1,4], Ahmed Abdelm…

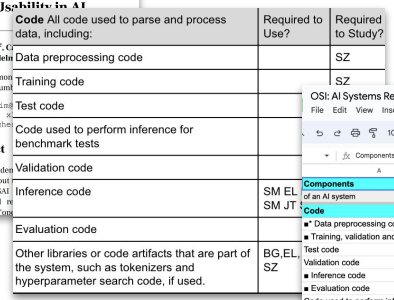
[1]LF AI & Data - Generative AI Common…
[3]University of Oxford, [4]Columb…

matt.white@berkeley.edu, ibrahim…
cailean.osborne@oii.ox.ac.uk, …
{ahmed.abdelmonsef,sachin.vargh…

Abstract

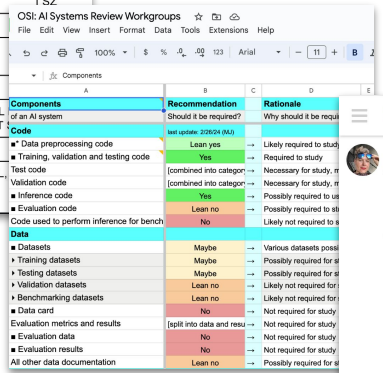
Generative AI (GAI) offers unpreceden…
commercialization has raised concerns about …
bias, and safety. Many "open-source" GAI …
components for full understanding and re…
restrictive licenses, a practice known as "ope…

Component Votes



	Required to Use?	Required to Study?
Code All code used to parse and process data, including:		
Data preprocessing code		SZ
Training code		SZ
Test code		
Code used to perform inference for benchmark tests		
Validation code		
Inference code		SM EL SM JT
Evaluation code		
Other libraries or code artifacts that are part of the system, such as tokenizers and hyperparameter search code, if used.		BG,EL SZ

Vote Compilation

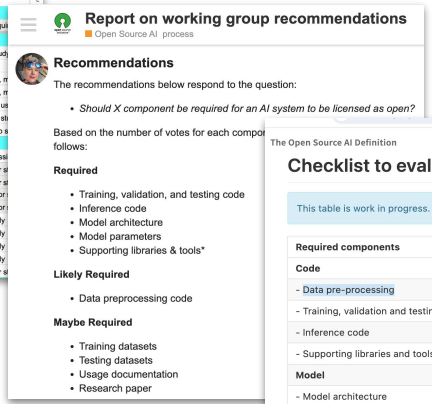


OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

Components	Recommendation	Rationale
of an AI system	Should it be required?	Why should it be requ…
Code	last update: 2/26/24 (MJ)	
▪ Data preprocessing code	Lean yes	Likely required to stud…
▪ Training, validation and testing code	Yes	Required to study
Test code	[combined into categor…	Necessary for study, r…
Validation code	[combined into categor…	Necessary for study, r…
▪ Inference code	Yes	Possibly required to us…
▪ Evaluation code	Lean no	Possibly required to st…
Code used to perform inference for bench…	No	Likely not required to s…
Data		
▪ Datasets	Maybe	Various datasets poss…
▸ Training datasets	Maybe	Possibly required for st…
▸ Testing datasets	Maybe	Possibly required for st…
▸ Validation datasets	Lean no	Likely not required for …
▸ Benchmarking datasets	Lean no	Likely not required for …
▪ Data card	No	Not required for study
Evaluation metrics and results	[split into data and resu…	Not required for study
▪ Evaluation data	No	Not required for study
▪ Evaluation results	No	Not required for study
All other data documentation	Lean no	Possibly required for st…

Recommendation Report



Report on working group recommendations

Open Source AI process

Recommendations

The recommendations below respond to the question:

- *Should X component be required for an AI system to be licensed as open?*

Based on the number of votes for each compo… follows:

Required

- Training, validation, and testing code
- Inference code
- Model architecture
- Model parameters
- Supporting libraries & tools*

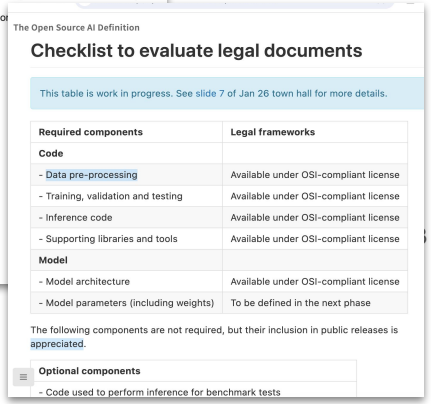
Likely Required

- Data preprocessing code

Maybe Required

- Training datasets
- Testing datasets
- Usage documentation
- Research paper

Definition Checklist



The Open Source AI Definition

Checklist to evaluate legal documents

This table is work in progress. See slide 7 of Jan 26 town hall for more details.

Required components	Legal frameworks
Code	
- Data pre-processing	Available under OSI-compliant license
- Training, validation and testing	Available under OSI-compliant license
- Inference code	Available under OSI-compliant license
- Supporting libraries and tools	Available under OSI-compliant license
Model	
- Model architecture	Available under OSI-compliant license
- Model parameters (including weights)	To be defined in the next phase

The following components are not required, but their inclusion in public releases is appreciated.

Optional components
- Code used to perform inference for benchmark tests

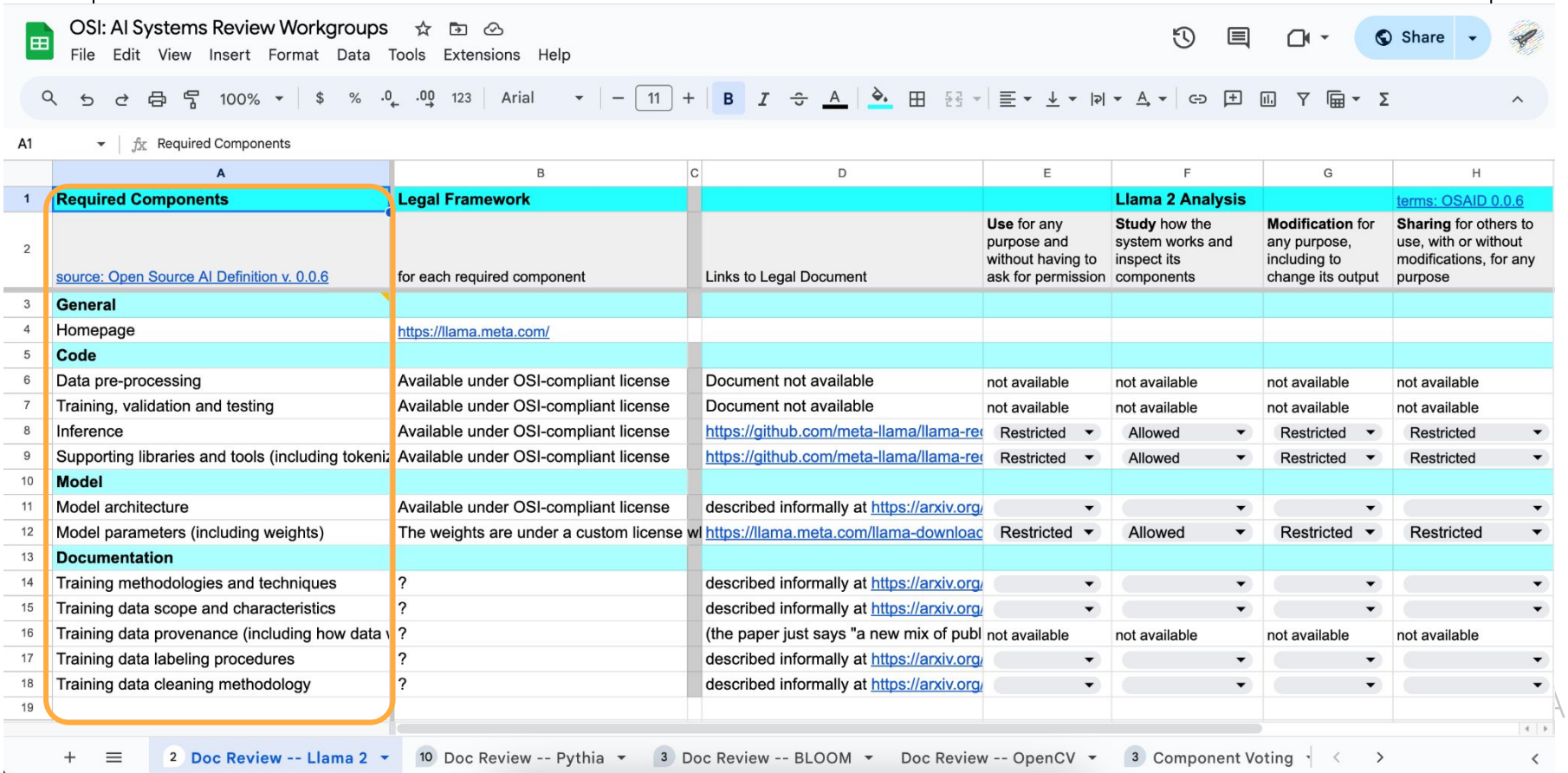
v.0.0.6

Open Source AI Definition

Legal Documents Review

Early Spring 2024

Workgroups: Legal Documents Review



OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

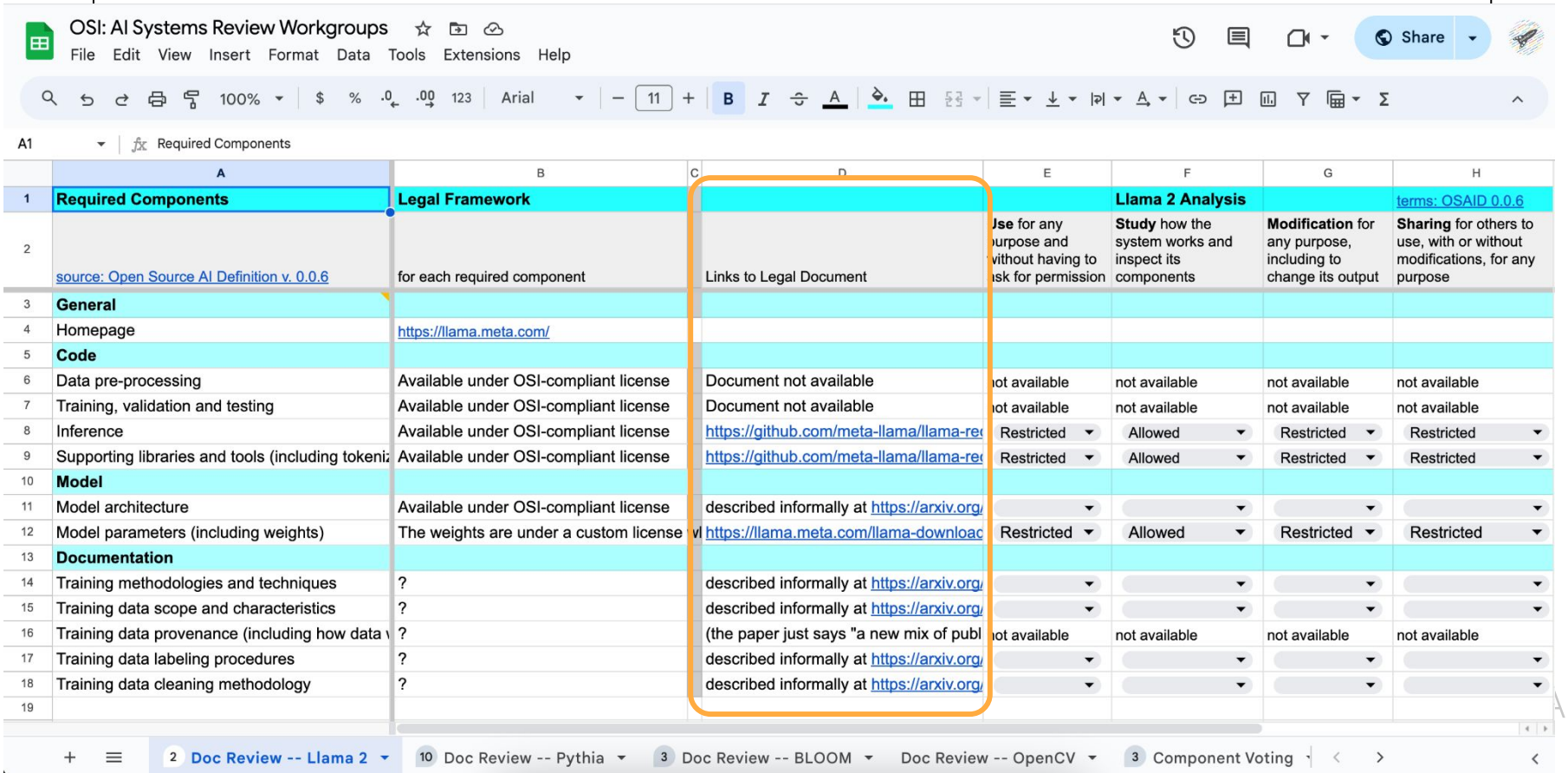
Share

A1 Required Components

	A	B	C	D	E	F	G	H
1	**Required Components**	**Legal Framework**				**Llama 2 Analysis**		terms: OSAID 0.0.6
2	source: Open Source AI Definition v. 0.0.6	for each required component		Links to Legal Document	**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	**Sharing** for others to use, with or without modifications, for any purpose
3	**General**							
4	Homepage	https://llama.meta.com/						
5	**Code**							
6	Data pre-processing	Available under OSI-compliant license		Document not available	not available	not available	not available	not available
7	Training, validation and testing	Available under OSI-compliant license		Document not available	not available	not available	not available	not available
8	Inference	Available under OSI-compliant license		https://github.com/meta-llama/llama-re	Restricted	Allowed	Restricted	Restricted
9	Supporting libraries and tools (including tokeni	Available under OSI-compliant license		https://github.com/meta-llama/llama-re	Restricted	Allowed	Restricted	Restricted
10	**Model**							
11	Model architecture	Available under OSI-compliant license		described informally at https://arxiv.org/				
12	Model parameters (including weights)	The weights are under a custom license wl		https://llama.meta.com/llama-downloac	Restricted	Allowed	Restricted	Restricted
13	**Documentation**							
14	Training methodologies and techniques	?		described informally at https://arxiv.org/				
15	Training data scope and characteristics	?		described informally at https://arxiv.org/				
16	Training data provenance (including how data v	?		(the paper just says "a new mix of publ	not available	not available	not available	not available
17	Training data labeling procedures	?		described informally at https://arxiv.org/				
18	Training data cleaning methodology	?		described informally at https://arxiv.org/				
19								

2 Doc Review -- Llama 2 | 10 Doc Review -- Pythia | 3 Doc Review -- BLOOM | Doc Review -- OpenCV | 3 Component Voting

Workgroups: Legal Documents Review



OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

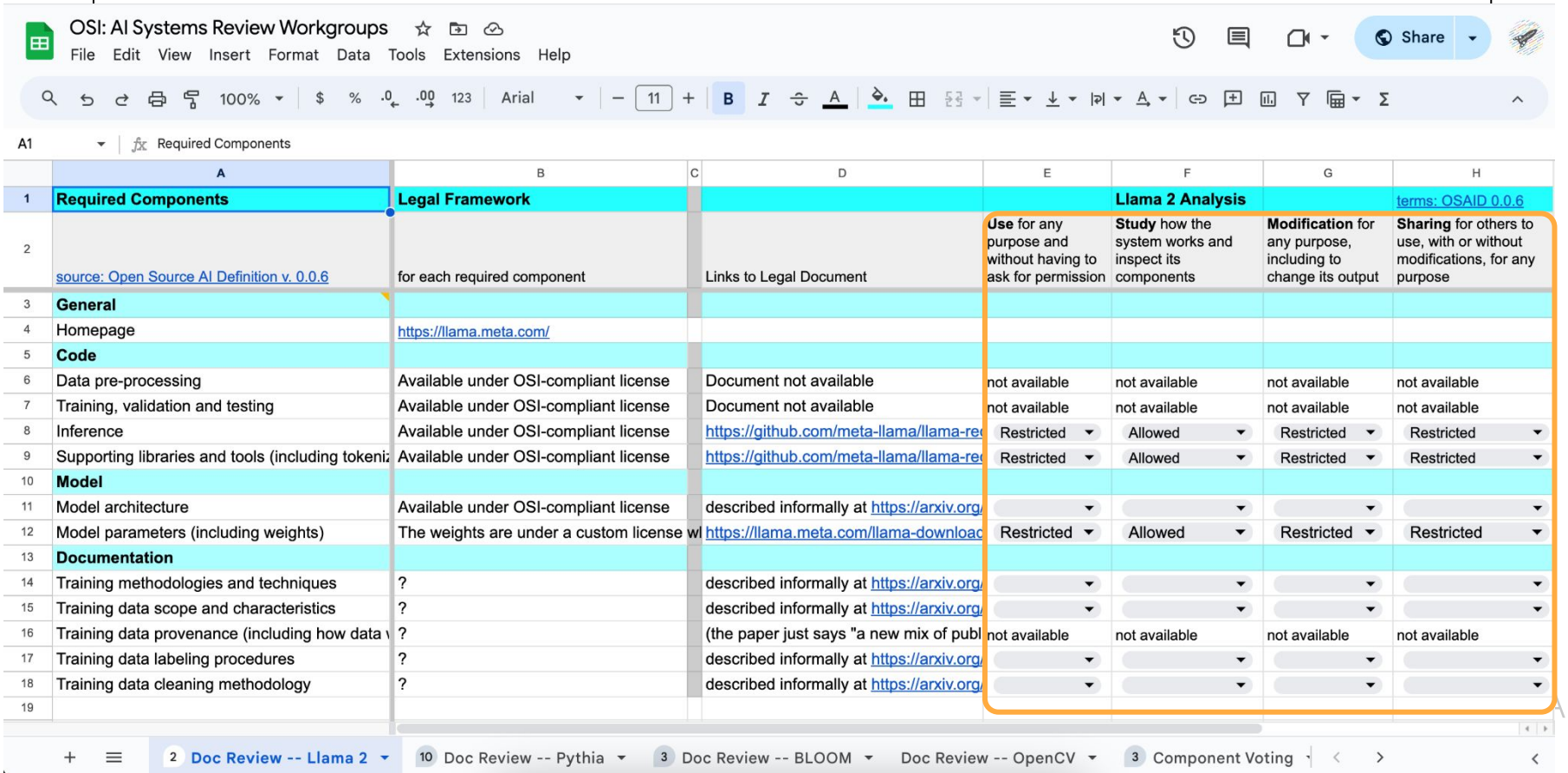
Share

A1 Required Components

	A	B	C	D	E	F	G	H
1	**Required Components**	**Legal Framework**				**Llama 2 Analysis**		terms: OSAID 0.0.6
2	source: Open Source AI Definition v. 0.0.6	for each required component		Links to Legal Document	**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	**Sharing** for others to use, with or without modifications, for any purpose
3	**General**							
4	Homepage	https://llama.meta.com/						
5	**Code**							
6	Data pre-processing	Available under OSI-compliant license		Document not available	ot available	not available	not available	not available
7	Training, validation and testing	Available under OSI-compliant license		Document not available	ot available	not available	not available	not available
8	Inference	Available under OSI-compliant license		https://github.com/meta-llama/llama-re	Restricted	Allowed	Restricted	Restricted
9	Supporting libraries and tools (including tokeni	Available under OSI-compliant license		https://github.com/meta-llama/llama-re	Restricted	Allowed	Restricted	Restricted
10	**Model**							
11	Model architecture	Available under OSI-compliant license		described informally at https://arxiv.org/				
12	Model parameters (including weights)	The weights are under a custom license vl		https://llama.meta.com/llama-downloac	Restricted	Allowed	Restricted	Restricted
13	**Documentation**							
14	Training methodologies and techniques	?		described informally at https://arxiv.org/				
15	Training data scope and characteristics	?		described informally at https://arxiv.org/				
16	Training data provenance (including how data v	?		(the paper just says "a new mix of publ	ot available	not available	not available	not available
17	Training data labeling procedures	?		described informally at https://arxiv.org/				
18	Training data cleaning methodology	?		described informally at https://arxiv.org/				
19								

2 Doc Review -- Llama 2 | 10 Doc Review -- Pythia | 3 Doc Review -- BLOOM | Doc Review -- OpenCV | 3 Component Voting

Workgroups: Legal Documents Review



OSI: AI Systems Review Workgroups

File Edit View Insert Format Data Tools Extensions Help

Share

A1 Required Components

	A	B	C	D	E	F	G	H
1	**Required Components**	**Legal Framework**				**Llama 2 Analysis**		terms: OSAID 0.0.6
2	source: Open Source AI Definition v. 0.0.6	for each required component		Links to Legal Document	**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	**Sharing** for others to use, with or without modifications, for any purpose
3	**General**							
4	Homepage	https://llama.meta.com/						
5	**Code**							
6	Data pre-processing	Available under OSI-compliant license		Document not available	not available	not available	not available	not available
7	Training, validation and testing	Available under OSI-compliant license		Document not available	not available	not available	not available	not available
8	Inference	Available under OSI-compliant license		https://github.com/meta-llama/llama-re	Restricted	Allowed	Restricted	Restricted
9	Supporting libraries and tools (including tokeni	Available under OSI-compliant license		https://github.com/meta-llama/llama-re	Restricted	Allowed	Restricted	Restricted
10	**Model**							
11	Model architecture	Available under OSI-compliant license		described informally at https://arxiv.org/				
12	Model parameters (including weights)	The weights are under a custom license wl		https://llama.meta.com/llama-downloac	Restricted	Allowed	Restricted	Restricted
13	**Documentation**							
14	Training methodologies and techniques	?		described informally at https://arxiv.org/				
15	Training data scope and characteristics	?		described informally at https://arxiv.org/				
16	Training data provenance (including how data v	?		(the paper just says "a new mix of publ	not available	not available	not available	not available
17	Training data labeling procedures	?		described informally at https://arxiv.org/				
18	Training data cleaning methodology	?		described informally at https://arxiv.org/				
19								

2 Doc Review -- Llama 2 | 10 Doc Review -- Pythia | 3 Doc Review -- BLOOM | Doc Review -- OpenCV | 3 Component Voting

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-

Open Source AI Definition **Required Components** v.0.0.8

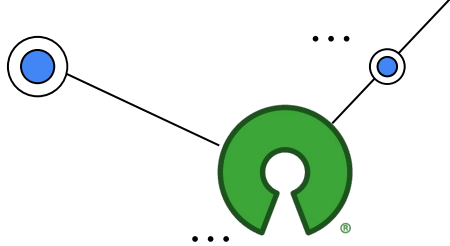
Optional components	Legal frameworks
Data information All data sets, including:	
- Training data sets	Available under OSD-compliant license
- Testing data sets	Available under OSD-compliant license
- Validation data sets	Available under OSD-compliant license
- Benchmarking data sets	Available under OSD-compliant license
- Data card	Available under OSD-compliant license
- Evaluation data	Available under OSD-compliant license
- Evaluation results	Available under OSD-compliant license
- Other data documentation	Available under OSD-compliant license
Code	
- Code used to perform inference for benchmark tests	Available under OSI-approved license
- Evaluation code	Available under OSI-approved license
Model All model elements, including:	
- Model card	Available under OSD-compliant license
- Sample model outputs	Available under OSD-compliant license
- Model metadata	Available under OSD-compliant license
Other Any other documentation or tools produced or used, including:	
- Research papers	Available under OSD-compliant license
- Technical report	Available under OSD-compliant

Open Source AI Definition **Optional Components** v.0.0.8

Open Source AI Definition

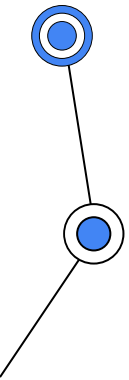
A Representative Process

Diversity, Inclusion, and Equity



Equitable and inclusive stakeholder representation isn't only about for justice, it's about legitimacy.

A global definition requires global consultation.



Stakeholder	Description	Example
1. **System Creator**	Makes AI system and/or component that will be studied, used, modified, or shared through an open source license	ML researcher in academia or industry
2. **License Creator**	Writes or edits the open source license to be applied to the AI system or component, includes compliance	IP lawyer
3. **Regulator**	Writes or edits rules governing licenses and systems	government policy-maker
4. **Licensee**	Seeks to study, use modify, or share an open source AI system	AI engineer in industry, health researcher in academia
5. **End User**	Consumes a system output, but does not seek to study, use, modify, or share the system	student using a chatbot to write a report, artist creating an image
6. **Subject**	Affected upstream or downstream by a system output without interacting with it intentionally + advocates for this group.	photographer who finds their image in training dataset (upstream), mortgage applicant evaluated by a bank's AI system (downstream)

Stakeholder	Description	Example
1. **System Creator**	Makes AI system and/or component that will be studied, used, modified, or shared through an open source license	ML researcher in academia or industry
2. **License Creator**	Writes or edits the open source license to be applied to the AI system or component, includes compliance	IP lawyer
3. **Regulator**	Writes or edits rules governing licenses and systems	government policy-maker
4. **Licensee**	Seeks to study, use modify, or share an open source AI system	AI engineer in industry, health researcher in academia
5. **End User**	Consumes a system output, but does not seek to study, use, modify, or share the system	student using a chatbot to write a report, artist creating an image
6. **Subject**	Affected upstream or downstream by a system output without interacting with it intentionally + advocates for this group.	photographer who finds their image in training dataset (upstream), mortgage applicant evaluated by a bank's AI system (downstream)

Most involved in current phase

open source initiative®

Seeking document reviewers for Pythia and OpenCV

Open Source AI process

Mer 5d

TASK: As part of the systems review track, we're looking for volunteers to review licenses for the Pythia and OpenCV systems and fill out this spreadsheet 4 to check the compatibility of version 0.0.6 3 of our definition with current AI systems.

TIMELINE: Our goal was to complete this review by next Tuesday, April 2nd, though we'll likely extend the deadline in consultation with the volunteers who respond.

VOLUNTEERS: For transparency, reviewers will have their names and affiliations made public. Black, Indigenous, Latine, and other people of color, women, queer, transgender, and non-binary people, people with disabilities, and people from poor and working class backgrounds are encouraged to respond.

LEARN MORE Reviewers are already assigned in the Llama 2 and BLOOM groups. We have two reviewers for Pythia and are seeking more. We have no reviewers yet for OpenCV. Further information on the workgroups and their past activities can be found here 2.

Open Source AI Definition
Next Steps
Spring - Fall, 2024

Definition Validation

- Confirm current systems (Llama 2, Pythia, BLOOM, OpenCV) equally reviewable under v. 0.0.8
- Seeking **volunteers** to review 1 to 3 additional AI systems to see how well they align with the definition
  - Contact Mer at Mer@dobiggood.com if you are interested.
- Due Monday, May 20th

Reviewers

We are interested in reviewing about 10 AI systems self-described as open as part of this definition process. Those marked (*) have were reviewed in previous phases. Other systems are newly added.

1. Arctic

1. Seeking volunteer

2. BLOOM*

2. **Danish Contractor** BLOOM Model Gov. Work Group
3. **Jaan Li** University of Tartu, Phare Health

3. Falcon

1. **Casey Valk** Nutanix

4. Grok

1. **Victor Lu** independent database consultant

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

7. OLMo

6. **Amanda Casari** Google
7. **Abdoulaye Diack** Google

8. OpenCV*

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

9. Phi-2

1. **Seo-Young Isabelle Hwang** Samsung
2. **Abdoulaye Diack** Google

10. Pythia*

3. **Seo-Young Isabelle Hwang** Samsung
4. **Stella Biderman** EleutherAI
5. **Hailey Schoelkopf** EleutherAI
6. **Aviya Skowron** EleutherAI

11. T5

1. **Jaan Li** University of Tartu, Phare Health

Additional volunteers welcome on all systems

2024 Timeline

System testing work stream

Stakeholder consultation work stream

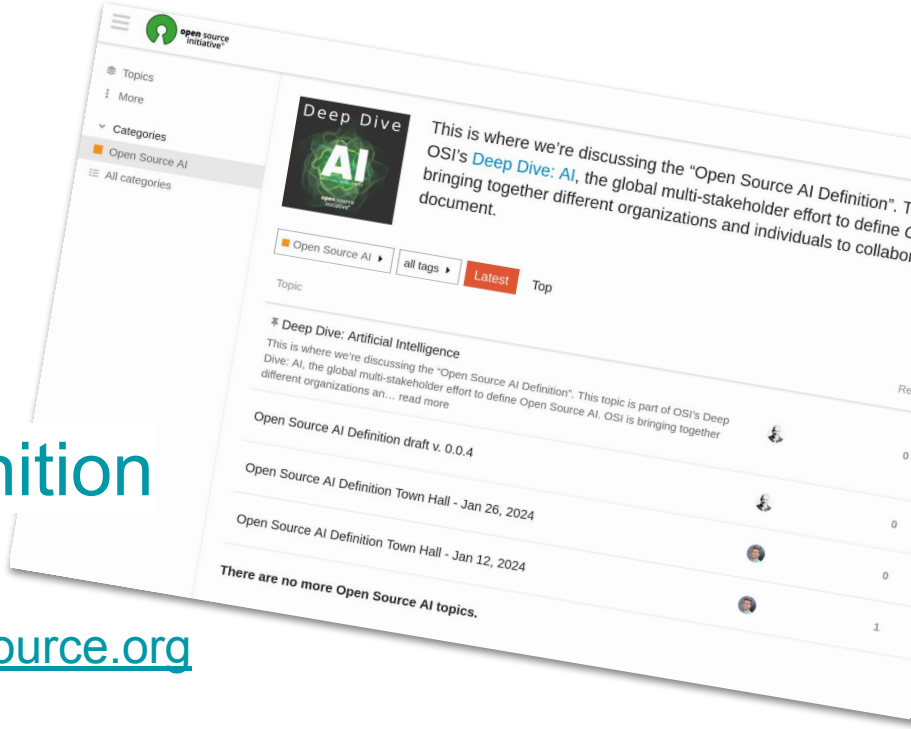
Release schedule

OSAID v. 0.0.8 last week

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)	September
North America	United States	Raleigh	**All Things Open**	Oct 27 - 29



Open Source AI Definition
Get Involved

- Public forum: discuss.opensource.org
- Become an OSI Member
  - Free or or full
  - SSO with other OSI websites
- Biweekly virtual town halls… like this one!
- **Volunteer** for definition validation (email Mer)

Q & A

Thank you

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