Opening up ChatGPT
a case study in operationalizing openness in AI

Andreas Liesenfeld & Mark Dingemanse
Centre for Language Studies, Radboud University
"Open source will be a cornerstone of Germany’s digital state."

"Open source as a critical component of scientific research"

"Open source by default" principle
Our starting question:
“open source” large language models are on the rise — but how open are they?
I. Peer-reviewed paper


II. Crowd-sourced live tracker
Surveying “openness” in ChatGPT-like text generators

- in complex AI systems, openness is never all-or-nothing
- our approach: decompose into relevant constituent elements

<table>
<thead>
<tr>
<th>Availability</th>
<th>Documentation</th>
<th>User access</th>
</tr>
</thead>
</table>

Surveying “openness” in ChatGPT-like text generators

- in complex AI systems, openness is never all-or-nothing
- our approach: decompose into relevant constituent elements
- for each element, record degree of openness

<table>
<thead>
<tr>
<th>Availability</th>
<th>Documentation</th>
<th>User access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open code</td>
<td>Code</td>
<td>Package</td>
</tr>
<tr>
<td>Base model data</td>
<td>Architecture</td>
<td>API</td>
</tr>
<tr>
<td>Base model weights</td>
<td>Preprint</td>
<td></td>
</tr>
<tr>
<td>RLHF data</td>
<td>Paper</td>
<td></td>
</tr>
<tr>
<td>RLHF weights</td>
<td>Model card</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td>Data sheet</td>
<td></td>
</tr>
<tr>
<td>Project</td>
<td>Availability</td>
<td>Documentation</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>BLOOMZ</td>
<td>bigscience-workshop</td>
<td>LLM base: BLOOMZ, mT0</td>
</tr>
<tr>
<td>Pythia-Chat-Base-78-1</td>
<td>togetherputer</td>
<td>LLM base: EleutherAI pythia</td>
</tr>
<tr>
<td>Open Assistant</td>
<td>LAION-AI</td>
<td>LLM base: Pythia 12B</td>
</tr>
<tr>
<td>dolly</td>
<td>databricks</td>
<td>LLM base: EleutherAI pythia</td>
</tr>
<tr>
<td>RedPajama-INCITE-78-1</td>
<td>TogetherComputer</td>
<td>LLM base: RedPajama-INCITE-78-Base</td>
</tr>
<tr>
<td>trtx</td>
<td></td>
<td>LLM base: various (pythia, flan, OPT)</td>
</tr>
<tr>
<td>MPT-7B Instruct</td>
<td>MosaicML</td>
<td>LLM base: MosaicML</td>
</tr>
<tr>
<td>MPT-30B Instruct</td>
<td></td>
<td>LLM base: MosaicML</td>
</tr>
<tr>
<td>Vicuna 13B v 1.3</td>
<td>LMSYS</td>
<td>LLM base: LLama</td>
</tr>
<tr>
<td>minChatGPT</td>
<td>Letharyanjielli</td>
<td>LLM base: GPT2</td>
</tr>
<tr>
<td>ChatRWKV</td>
<td>BlinkDL/RWKV</td>
<td>LLM base: RWKV-LM</td>
</tr>
<tr>
<td>OpenChat V3</td>
<td>OpenChat</td>
<td>LLM base: Llama2</td>
</tr>
</tbody>
</table>
Introducing The World’s Largest Open Multilingual Language Model: BLOOM

Large language models (LLMs) have made a significant impact on AI research. These powerful, general models can take on a wide variety of new language tasks from a user’s instructions. However, academia, nonprofits, and even smaller companies’ research labs find it difficult to create, study, or even use LLMs as only a few industrial labs with the necessary resources and exclusive rights can fully access them. Today, we release BLOOM, the first multilingual LLM trained in complete transparency, to change this status quo — the result of the largest collaboration of AI researchers ever involved in a single research project.

Meet BLOOMChat: An Open-Source 176-Billion-Parameter Multilingual Chat Large Language Model (LLM) Built on Top of the BLOOM Model

How to use this table. Every cell records a three-level openness judgement (open, partial, or closed) with a direct link to the available evidence; on hover, the cell will display the notes we have on file for that judgement. At the end of a row, the § is a direct link to source data. The table is sorted by cumulative openness, where open is 1, ~ is 0.5 and closed is 0 points.
**Llama2** by Meta Platforms, Inc.

### Introducing Llama 2
The next generation of our open source large language model
Llama 2 is available for free for research and commercial use.

**Takeaways**
- Today, we’re introducing the availability of Llama 2, the next generation of our open source large language model.
- Llama 2 is free for research and commercial use.

### Meta’s Open Source Llama Upsets the AI Horse Race

<table>
<thead>
<tr>
<th>Project</th>
<th>Availability</th>
<th>Documentation</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook Research</td>
<td><img src="check" alt="" /> LLM base: LLaMA2</td>
<td><img src="check" alt="" /> RL base: Meta, StackExchange, Anthr...</td>
<td><img src="check" alt="" /></td>
</tr>
</tbody>
</table>

*How to use this table.* Every cell records a three-level openness judgement ( *[open](#) or *- [partial](#) or *- [closed](#) *) with a direct link to the available evidence; on hover, the cell will display the notes we have on file for that judgement. At the end of a row, the § is a direct link to source data. The table is sorted by cumulative openness, where *- is 1, *- is 0.5 and *- is 0 points.*
<table>
<thead>
<tr>
<th>Open code</th>
<th>No training code available</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLM data</td>
<td></td>
</tr>
<tr>
<td>LLM weights</td>
<td></td>
</tr>
<tr>
<td>RLHF data</td>
<td></td>
</tr>
<tr>
<td>RLHF weights</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td></td>
</tr>
<tr>
<td>Architecture</td>
<td></td>
</tr>
<tr>
<td>Preprint</td>
<td></td>
</tr>
<tr>
<td>Paper</td>
<td></td>
</tr>
<tr>
<td>Modelcard</td>
<td></td>
</tr>
<tr>
<td>Datasheet</td>
<td></td>
</tr>
<tr>
<td>Package</td>
<td></td>
</tr>
<tr>
<td>API</td>
<td></td>
</tr>
</tbody>
</table>

**Is the source code openly available?**

Fully shared code to enable reproduction
Is the pretraining dataset documented and available?
Fully shared code to enable reproduction  
Training data shared  
Access to base LLM without instruction tuning  

Open code  
LLM data  
**LLM weights**  
RLHF data  
RLHF weights  
License  
Code  
Architecture  
Preprint  
Paper  
Modelcard  
Datasheet  
Package  
API  

Are the model weights openly available?
<table>
<thead>
<tr>
<th>Open code</th>
<th>LLM data</th>
<th>LLM weights</th>
<th>RLHF data</th>
<th>RLHF weights</th>
<th>License</th>
<th>Code</th>
<th>Architecture</th>
<th>Preprint</th>
<th>Paper</th>
<th>Modelcard</th>
<th>Datasheet</th>
<th>Package</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessible</td>
<td>Accessible after registration</td>
<td>No training code available</td>
<td>No training data available</td>
<td>Accessible after registration</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Are the instruction-tuning datasets documented and available?
Are the instruction-tuned model weights made available?
Fully shared code to enable reproduction
Training data shared
Access to base LLM without instruction tuning
Accessible
Training checkpoint available to download
Code: Apache 2, Model: “RAIL”, non OSI

Open code
LLM data
LLM weights
RLHF data
RLHF weights
License
Code
Architecture
Preprint
Paper
Modelcard
Datasheet
Package
API

Is the system released under an open license?

No training code available
No training data available
Accessible after registration
No data
Accessible after registration
“Community license”, non OSI
Is the system released under an open license?

- Fully shared code to enable reproduction
- Training data shared
- Access to base LLM without instruction tuning

Accessible
- Training checkpoint available to download

Code: Apache 2, Model: “RAIL”, non OSI

Is “unlimited” always best?
- Responsible AI License (RAIL) aims to address the moral dilemmas of harmful and unintended uses of tech (Contractor et al. 2022 FAccT)
- Restricts particular use cases (e.g. “don’t use to exploit vulnerabilities of a specific group”)

Responsibility: two approaches
- Llama2: you may not “represent that Llama 2 outputs are human-generated” (a low bar)
- RAIL: you may not “generate content without expressly and intelligibly disclaiming that the text is machine generated”
Fully shared code to enable reproduction
Training data shared
Access to base LLM without instruction tuning
Accessible
Training checkpoint available to download
Code: Apache 2, Model: “RAIL”, non OSI
Accessible and well-maintained

Open code
LLM data
LLM weights
RLHF data
RLHF weights
License
Code
Architecture
Preprint
Paper
Modelcard
Datasheet
Package
API

No training code available
No training data available
Accessible after registration
No data
Accessible after registration
“Community license”, non OSI
Only minimal examples

Is the codebase well-maintained and documented?
Fully shared code to enable reproduction
Training data shared
Access to base LLM without instruction tuning
Accessible
Training checkpoint available to download
Code: Apache 2, Model: “RAIL”, non OSI
Accessible and well-maintained
Accessible and documented in preprint

Open code
LLM data
LLM weights
RLHF data
RLHF weights
License
Code

Architecture
Preprint
Paper
Modelcard
Datasheet
Package
API

No training code available
No training data available
Accessible after registration
No data
Accessible after registration
“Community license”, non OSI
Only minimal examples
Sketched in preprint

Is the system architecture clearly documented?
Fully shared code to enable reproduction
Training data shared
Access to base LLM without instruction tuning
Accessible
Training checkpoint available to download
Code: Apache 2, Model: “RAIL”, non OSI
Accessible and well-maintained
Accessible and documented in preprint
Multiple detailed preprints

Open code
LLM data
LLM weights
RLHF data
RLHF weights
License
Code
Architecture
Preprint
Paper
Modelcard
Datasheet
Package
API

No training code available
No training data available
Accessible after registration
No data
Accessible after registration
“Community license”, non OSI
Only minimal examples
Sketched in preprint
Corporate preprint only

Is there a preprint providing scientific documentation of the system?
Has the system been scrutinized under rigorous peer-review?
Is the model described in a model card? (Mitchell et al. 2019)
Fully shared code to enable reproduction
Training data shared
Access to base LLM without instruction tuning
Accessible
Training checkpoint available to download
Code: Apache 2, Model: “RAIL”, non OSI
Accessible and well-maintained
Accessible and documented in preprint
Multiple detailed preprints
No paper
Available
Available

Open code
LLM data
LLM weights
RLHF data
RLHF weights
License
Code
Architecture
Preprint
Paper
Modelcard
Datasheet
Package
API

No training code available
No training data available
Accessible after registration
No data
Accessible after registration
“Community license”, non OSI
Only minimal examples
Sketched in preprint
Corporate preprint only
No paper
Only minimal detail provided
No datasheet

Is there a data sheet documenting data collection & curation? (McMillan Major et al. 2023)
Is there a packaged release available?
Fully shared code to enable reproduction
Training data shared
Access to base LLM without instruction tuning
Accessible
Training checkpoint available to download
Code: Apache 2, Model: “RAIL”, non OSI
Accessible and well-maintained
Accessible and documented in preprint
Multiple detailed preprints
No paper
Available
Available
No package
“Petals API” available via huggingface

Open code
LLM data
LLM weights
RLHF data
RLHF weights
License
Code
Architecture
Preprint
Paper
Modelcard
Datasheet
Package
API

No training code available
No training data available
Accessible after registration
No data
Accessible after registration
“Community license”, non OSI
Only minimal examples
Sketched in preprint
Corporate preprint only
No paper
Only minimal detail provided
No datasheet
No package
Limited access, sign-up required

Is there an openly available API with unrestricted access?
Two extremes

- Both claim to be “open source” — only one is
- Drilling into details makes differences visible
- Evidence-based judgements help
  - to credit initiatives for care taken in developing and releasing AI technology
  - to puncture corporate hype
  - to call out hijacking of terms like “open source”

<table>
<thead>
<tr>
<th>Availability</th>
<th>Documentation</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open code</td>
<td>Code</td>
<td>Package</td>
</tr>
<tr>
<td>LLM data</td>
<td>Architecture</td>
<td>API</td>
</tr>
<tr>
<td>LLM weights</td>
<td>Preprint</td>
<td></td>
</tr>
<tr>
<td>RLHF data</td>
<td>Paper</td>
<td></td>
</tr>
<tr>
<td>RLHF weights</td>
<td>Modelcard</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td>Datasheet</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Package</td>
<td></td>
</tr>
<tr>
<td></td>
<td>API</td>
<td></td>
</tr>
</tbody>
</table>
Surveying 25+ text generators: recurring issues

1. Inherited data is common & legal murkiness ensues
2. Synthetic data is on the rise, with unknown consequences
3. “Release by blogpost” should not be accepted as sufficient

>40% of LLMs we survey now use *synthetic data* for instruction-tuning

* prompts, responses, or ratings harvested from other LLMs
• Current AI systems are complex and multi-part
Current AI systems are complex and multi-part — how to reverse engineer?
Downstream elements can obstruct access to earlier parts.
Current AI systems are complex and multi-part — how to reverse engineer?

Downstream elements can **obstruct** access to earlier parts ("roadblocks")

True openness only possible if intermediate steps **documented & opened up**

Supply source at each roadblock to preserve reverse engineerability
Conclusions

Our approach
- Isolate most relevant dimensions of openness (relative to system)
- Provide evidence-based judgements of openness on those
- All work done out in the open: opening-up-chatgpt.io

Towards a definition of “open” AI systems
- For any genAI system, openness will be composite & graded
- No one-size fits all solution: domain knowledge needed to identify relevant dimensions
- Preserve the spirit of reverse engineerability