5 Ways Open Source is Critical to the Future of AI

February 7, 2025

National Open Source Innovation Summit Dublin

Sayeed Choudhury – @sayeedc.bsky.social
Associate Dean for Digital Infrastructure
Director of Open Source Programs Office (OSPO)
Executive Director of Open Forum for AI (OFAI)

Open Forum for AI (OFAI)

CMU Launches New Initiative for Human-Centered AI

open source
initiative®

What's Open Source AI?

Following the same idea behind Open Source Software,
an Open Source AI is a system made available under terms that grant users the freedoms to:

Use   Study   Modify   Share

Use the system for any purpose and without having to ask for permission.

Precondition to exercise these freedoms is to have access to
the preferred form to make modifications to the system, and to the means to use it.

Carnegie Mellon University

Openness in AI Framework

- Technical framework, developed with community engagement, that rigorously defines openness and the ensuing or corresponding value

- Move from open source to open governance (e.g., transparency, reproducibility, accessibility)

- Need to consider specific context and use case with emphasis on the Ws of Why, What, Who, Whom, When, and Where

5 Ways…

- Building on successes of open source software

- Openness as a design principle (e.g., TCP/IP, HTTP)

- Pathway for open infrastructure and broader participation

- Enabler of addressing market failures

- Engagement with global South (aka global majority)

5 Ways…

- **Building on successes of open source software**

- Openness as a design principle (e.g., TCP/IP, HTTP)

- Pathway for open infrastructure and broader participation

- Enabler of addressing market failures

- Engagement with global South (aka global majority)

The Value of Open Source Software

Manuel Hoffmann[a]   Frank Nagle[b]   Yanuo Zhou[c]

This version: January 1, 2024

Abstract

The value of a non-pecuniary (free) product is inherently difficult to assess. A pervasive example is open source software (OSS), a global public good that plays a vital role in the economy and is foundational for most technology we use today. However, it is difficult to measure the value of OSS due to its non-pecuniary nature and lack of centralized usage tracking. Therefore, OSS remains largely unaccounted for in economic measures. Although prior studies have estimated the supply-side costs to recreate this software, a lack of data has hampered estimating the much larger demand-side (usage) value created by OSS. Therefore, to understand the complete economic and social value of widely-used OSS, we leverage unique global data from two complementary sources capturing OSS usage by millions of global firms. We first estimate the supply-side value by calculating the cost to recreate the most widely used OSS once. We then calculate the demand-side value based on a replacement value for each firm that uses the software and would need to build it internally if OSS did not exist. We estimate the supply-side value of widely-used OSS is $4.15 billion, but that the demand-side value is much larger at $8.8 trillion. We find that firms would need to spend 3.5 times more on software than they currently do if OSS did not exist. The top six programming languages in our sample comprise 84% of the demand-side value of OSS. Further, 96% of the demand-side value is created by only 5% of OSS developers.

Carnegie Mellon University

5 Ways…

- Building on successes of open source software

- **Openness as a design principle (e.g., TCP/IP, HTTP)**

- Pathway for open infrastructure and broader participation

- Enabler of addressing market failures

- Engagement with global South (aka gl

The growing focus on open source AI makes it all the more important that we establish a shared understanding of what open source AI is. A definition should outline what must be shared and under what terms or conditions. Without this clarity, we risk a fragmented approach, where companies label their products as "open source" even when they aren't, where civil society doesn't have access to the AI components they need for testing and accountability, and policymakers create regulations that fail to address the complexities of the issue.

The Open Source Initiative (OSI) has recently released a new draft definition of open source AI, marking a critical juncture in the evolution of the internet. This moment comes after two years of conversations, debates, engagements, and late-night conversations across the technical and open source communities. It is critical not just for redefining what "open source" means in the context of AI; it's about shaping the future of the technology and its impact on society.

5 Ways…

- Building on successes of open source software

- Openness as a design principle (e.g., TCP/IP, HTTP)

- **Pathway for open infrastructure and broader participation**

- Enabler of addressing market failures

- Engagement with global South (aka gl

deepseek

Into the unknown

Start Now

Free access to DeepSeek-V3.

Experience the intelligent model.

Get DeepSeek App

Chat on the go with DeepSeek-V3

Your free all-in-one AI tool

5 Ways…

- Building on successes of open source software

- Openness as a design principle (e.g., TCP/IP, HTTP)

- Pathway for open infrastructure and broader participation

- **Enabler of addressing market failures**

- Engagement with global South (aka global majority)

Together, there's something science can do.

Too often, patients and families receive a difficult diagnosis, followed by "there's nothing we can do." On our own, that might be true, but by treating science as a team sport, using AI and open science to advance drug discovery, we see a world where medical advances leave no one behind.

About us

5 Ways…

- Building on successes of open source software

- Openness as a design principle (e.g., TCP/IP, HTTP)

- Pathway for open infrastructure and broader participation

- Enabler of addressing market failures

- **Engagement with global South (aka global majority)**

Pangea: A Fully Open *Multilingual Multimodal* LLM for 39 Languages

Pangea-7B: a strong multilingual multimodal LLM capable of 39 languages.

PangeaIns: a 6M multilingual multimodal instruction tuning dataset spanning 39 languages.

PangeaBench: a holistic evaluation benchmark spanning 14 datasets in 47 languages.

arXiv   PDF   Code   Checkpoints

PangeaInstruct   PangeaBench   Demo

Xiang Yue[*]   Yueqi Song[*]   Akari Asai   Seungone Kim

Jean de Dieu Nyandwi   Simran Khanuja   Anjali Kantharuban

Lintang Sutawika   Sathyanarayanan Ramamoorthy   Graham Neubig

Carnegie Mellon University

Carnegie Mellon University

Acknowledgements and Q&A