If open source is to win, it must go public

Joshua Tan [1 2] **Nicholas Vincent** [3 1] **Katherine Elkins** [4] **Magnus Sahlgren** [5]

Abstract

Open source projects have made incredible progress in producing transparent and widely usable machine learning models and systems, but open source alone will face challenges in fully democratizing access to AI. Unlike software, AI models require substantial resources for activation—compute, post-training, deployment, and oversight—which only a few actors can currently provide. This paper argues that open source AI must be complemented by public AI: infrastructure and institutions that ensure models are accessible, sustainable, and governed in the public interest. To achieve the full promise of AI models as prosocial public goods, we need to build public infrastructure to power and deliver open source software and models.

1. Introduction

Open source, and the ethos of openness, has long served as a counterweight to concentrated control in computing. From Linux to Kubernetes, open collaboration has enabled researchers, companies, and the public at large to build on shared and trustworthy infrastructure (Raymond, 1999; Eghbal, 2016). But open source has always straddled a line between the emancipatory ideals of the Free Software movement and the strategic goals of firms (Weber, 2005; Kelty, 2008). What appears as a spontaneous gift economy is often scaffolded by sponsorships, employment arrangements, and various private (and public) subsidies. This compromise between community and commerce has proven remarkably productive in many software categories like cloud computing, programming languages, and operating systems. But it's breaking down for the largest foundation models in AI. Such large language models (LLMs) are incredibly expensive to train, raising questions around the longevity of open models (Maslej et al., 2025; Choksi et al., 2025). Once trained, open source weights alone are inert; without inference, fine-tuning, localization, tooling, and interfaces, they remain unusable to all but a small elite with the capital, compute, and engineering to deploy them (Bommasani et al., 2024; HAI, 2024). And even if deployed, the decentralized nature of open source deployments means that important RLHF and query data can become stranded across many silos. As modern AI ecosystems mature, so too must our expectations about what open source practices can–and cannot–deliver for researchers, for firms, and for the public.

In this paper, we argue that without structural intervention from public institutions, current open source efforts in AI will not democratize access to AI nor provision public goods (in the technical sense of non-rivalrous and non-excludable goods) as comparable open source efforts have done in other categories. This will hurt the machine learning research community. It will hurt startups. It will also undermine the strategic interests of large firms promoting open weight and open source AI. To move forward, we need to build broader public AI ecosystems that ensure open source AI is accessible, trustworthy, and competitive with closed source alternatives.

2. Background on open source AI

Here, we briefly summarize open source software in ML and the role of open source AI projects in proving the viability of openness. This is not meant to comprehensively cover all the open software that plays a role in the ML research stack nor cover all successful open source AI projects, of which there have been many.

Open Source in ML: The machine learning community has embraced open source as both a cultural and technical norm. Frameworks like PyTorch (Paszke et al., 2019), Hugging Face Transformers (Wolf et al., 2019), and Diffusers (von Platen et al., 2022) have made advanced ML tools widely accessible. Researchers routinely release code and models alongside publications, and open pre-trained weights have accelerated innovation and experimentation. Many foundational libraries like NumPy have long been released as open source software, which helped reduce reliance on proprietary scientific software. Relatedly, peer production platforms like Wikimedia projects (including Wikipedia)

[*]Equal contribution [1]Public AI Network [2]Metagov [3]Simon Fraser University, British Columbia, Canada [4]Kenyon College [5]AI Sweden. Correspondence to: Joshua Tan <josh@metagov.org>, Nicholas Vincent <nvincent@sfu.ca>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

have long been central to ML training data (Johnson et al., 2024).

Open Source AI Projects: Various open source community communities have been extremely successful in producing AI projects, covering general models, domain specific models, datasets, benchmarks, and more. EleutherAI's GPT-NeoX and later Pythia models (Black et al., 2022; Biderman et al., 2023) provide fully open GPT models that have been widely used for practical and scientific purposes. On the data front, the LAION project was critical for providing open data for computer vision (Schuhmann et al., 2022). A wide array of other projects, including RWKV (Peng et al., 2023), BigScience (BigScience Workshop et al., 2023), BigCode (Li et al., 2023), OpenFold (Ahdritz et al., 2024), RedPajamas-INCITE (Weber et al., 2024), OLMo (Groeneveld et al., 2024), and more recently Marin (Hall et al., 2025) have all contributed to an ecosystem of usable open source AI.

3. Challenges posed by AI models

However, AI models are not like traditional codebases.[1] The costs of training, running, and maintaining a frontier model vastly exceeds that of compiling and distributing software:

Pretraining Requires Capital and Scale: Modern models require substantial energy (Luccioni et al., 2024) and are trained on thousands of GPUs over weeks or months (Maslej et al., 2025). This demands not only access to ever-larger compute clusters–often only available to well-funded corporations or state-supported institutions, though progress is being made on distributed training–but also massive datasets, robust engineering teams, and complex distributed training infrastructure.

Inference is Not Free: Unlike a software library which users can potentially run using widely available CPUs, inference at scale (especially for large models) demands ongoing GPU access, orchestration systems, and cost management. These are nontrivial operational burdens that currently rely on a small number of hosting providers. While there is (sometimes) research funding for training, there is comparatively little research funding for inference—the National Deep Inference Fabric being one of the few exceptions (Fiotto-Kaufman et al., 2025)—with concomitant gaps in equitable access to state-of-the-art capabilities. As a result, even open-weight models often remain inert for most users, functionally gated behind commercial APIs or expensive compute setups.

Post-training involves proprietary data and design choices: The fine-tuning, alignment, tool integration, and prompt orchestration that make models actually useful in practice are often kept closed. While model weights may be public, the systems that give them utility are private.

Beyond the costs, open models face other problems:

Licensing is Ambiguous or Fragile: As an example, the Llama license, while widely used, contains restrictive terms and explicit revocability. Efforts to describe Llama as open source have been widely criticized by the OSS community, as the models released by Meta have not met the agreed standard for "open source" (OSI Opinion & Maris, 2025). Furthermore, companies like Meta can unilaterally stop releasing models or add even more restrictions to future licenses at any time.

Transparency is Partial and Inconsistent: While non-profit and academic models have maintained an admirable commitment to openness, the most powerful "open" models are trained by private companies and still largely fail to provide details on training data, compute budgets, or evaluation procedures.

Closed-source Co-optation: Open-source is built on a compromise between community and commerce, but community-contributed evaluations, tooling, datasets, and fine-tuning techniques often accrue value to large firms whose commitment to open source is tenuous at best—in this case the frontier labs who train closed-source models. Open-source contributors imagine they are building shared infrastructure; in reality, they may be fueling a pipeline that concentrates power (Widder et al., 2024).

The AI ecosystem is also evolving quickly. What began as token prediction from weights that fit on a local GPU is morphing into AI assistants that blend multi-modal reasoning, access to proprietary tools, and complex orchestration layers. As system complexity grows, the gap between "available weights" and "usable systems" widens. While there has been massive progress in terms of what can be run locally on consumer hardware (Gerganov, 2023; Ollama Team, 2025), these models still lack post-training alignment, retrieval augmentation, tool use integration, usage analytics, uptime guarantees, and continual updates that distinguish private,

[1] In practice, frontier models require private complements such as compute and energy to be useful. In economic terms, that means that they are "impure public goods" (Reiss, 2021; Eaves et al., 2024) or club goods (Gries & Naudé, 2022) rather than pure public goods. A classic example of an impure public good is a lighthouse financed by "light dues" (Reiss, 2021) imposed on ship owners. The eventual economic model for AI could end up looking like this, though the question of who collects the dues will be salient to whether access to AI is democratized. For an example that more closely captures AI models, we might imagine a library whose collection is non-rival and openly licensed. As the catalog grows by orders of magnitude, it becomes impossible for typical users to find a book without hiring a private "guide" to help. The books remain true public goods, but access to their knowledge is mediated by a search-and-retrieval toll good, so access becomes "club-like" and effectively the library provides information as an impure public good.

hosted services from merely downloadable weights. This is not a failure of values but of structure. The open source model, which flourished in an earlier era of low-cost computation and interoperable standards, is no longer sufficient on its own. To deliver on the promise of accessible and democratic AI (Collective Intelligence Project, 2024), we must build new public AI infrastructures that can provision and govern the full model lifecycle, not just its first checkpoint.

4. Position Statement

We assert: *open source AI, as currently practiced, will not by itself democratize access to AI or provision public goods as comparable open source efforts have done in other software categories.* Instead, we propose that open source AI must be embedded within a broader vision of public AI, defined by the following principles:

- Public Support: There must be public funding and infrastructure for inference, deployment, post-training, and data flywheels, not just pretraining.
- Public Access: Everyone—-global south researchers, civic technologists, local communities outside of Big Tech—-must be enabled to build, adapt, and use competitive models.
- Public Governance: Institutions accountable to the public—-governments, national labs, public utilities, universities, and nonprofits—must provision, host, and maintain models and related infrastructure.
- Private Commitments: Private actors must be encouraged (or required) to make commitments around openness, safety, and community control.

Public AI is not a theoretical aspiration. Around the world, countries are already experimenting with concrete strategies for building and deploying large-scale AI systems in the public interest. These approaches follow four broad models (Sitaraman & Parek, 2025):

- Outsourced Provision: Governments contract private labs to deliver models and compute, e.g. IndiaAI or the USA's National AI Research Resource (NAIRR) pilot.
- Networked Collaboration: Academic and civic actors co-develop systems, e.g. Empire AI in New York State or Canada's public compute investments.
- State-Corporate Fusion: Close governmental control over private corporations, e.g. China's AI programs or aspects of the UAE's AI strategy.
- Public Options: Direct public provision of AI services, e.g. Sweden's GPT-SW3 or Japan's ABCI.

New approaches are still being developed, including the recent Airbus for AI (Valero & Crespo, 2024; Tan et al., 2025) and CERN for AI (CAIRNE; Juijn et al., 2024) proposals for multilateral visions of public AI.

5. Alternative Views

We recognize a number of serious alternative views that challenge the necessity or feasibility of public AI.

5.1. View 1: The Market Is Working. Let OpenAI and Meta Lead.

Many believe that the private sector is successfully scaling AI access. OpenAI, Meta, Mistral, and DeepSeek have made advanced models available cheaply or for free. Proprietary labs have shown tremendous speed and capability in model iteration, evaluation, and deployment. Their models are at the performance frontier, their user experience is polished, and their costs are rapidly dropping.

Response: Access is not governance. These systems remain opaque and subject to unilateral revocation. The ability to use a chatbot today does not ensure access to trustworthy, auditable systems tomorrow. Public AI is not about replacing private labs, but about ensuring that there are durable, open, and accountable systems aligned with public needs and values. For example, the USA's National Deep Inference Fabric was designed to provide democratic access to open-weight models (Fiotto-Kaufman et al., 2025), while Sweden's GPT-SW3 (Ekgren et al., 2024) was initially trained to address ChatGPT's poor performance in Swedish and other Nordic languages.

It is also worth noting that Meta and other corporate-OSS builders stand to benefit if open source inference becomes publicly funded: the cheaper the inference, the more value accrues to the application layer.

5.2. View 2: Open Source Will Win Eventually. Just Be Patient.

This view argues that the open source ecosystem is improving rapidly (Maslej et al., 2025) and will eventually produce models on par with or better than proprietary models. The release of high-quality weights (e.g., Mistral, DeepSeek, Zephyr), coupled with open fine-tuning libraries and model merging techniques, suggests that community-driven innovation will outcompete closed models in the long run.

Response: Open source progress has been remarkable. There have been many academic, nonprofit, and community-led efforts to train foundation models. But all of the strongest and most-used open source models today were pre-trained by well-capitalized private companies: compare Llama 3.1-8B's 6M monthly downloads on HuggingFace with Eleuther Pythia's 900k and OLMo 2-7B's 29k (as of late May 2025) (Hugging Face, 2025). And none of these options compare with the (consumer) adoption seen by OpenAI and Anthropic—not to mention the potential for extraordinary adoption as proprietary models are incorporated into product platforms like Google Search or Microsoft Office. Public AI offers a mechanism to make open source development sustainable, pluralistic, and public-benefit-oriented. It ensures that open source models remain accessible, trustworthy, and responsive to broad public needs rather than to the incentives of a single commercial sponsor.

5.3. View 3: OSS + Hosting Already Works.

Why add bureaucracy? A practical open source ecosystem is already in place. Open models are hosted via Hugging Face, Replicate, and Open Router. Inference is affordable. User-facing products are emerging. Why burden this with new governance structures or public spending?

Response: Like view 1, this view confuses current availability with long-term stability. Most current deployments rely on ephemeral commercial hosting or terms that can be revoked. The fragility of the OSS+hosting stack is exemplified by the LLaMA license and the risk of unilateral pullback from companies like Meta. Public AI does not aim to replace this ecosystem but to underwrite it. This already happens, especially for academic and nonprofit projects: for example, national labs in the US use EleutherAI's GPT-NeoX and have provided some support for the project, while the French National Center for Scientific Research supported the BLOOM project, which was trained on a French public supercomputer (BigScience Workshop et al., 2023).

5.4. View 4: Regulation Is a Better Tool Than Public Investment.

Instead of building new infrastructure, governments can simply regulate AI development—imposing transparency requirements, safety standards, and licensing constraints. Regulatory frameworks such as the EU AI Act and export controls on GPU sales aim to shape the AI landscape through law rather than through investment.

Response: Regulation is essential, but it is not sufficient. It can curb harmful behavior but does little to guarantee access, usability, or equitable participation. Public AI is proactive: it builds capabilities and institutions that embody public values from the outset. Rather than rely solely on constraints imposed on private actors, public AI enables public-purpose development from the ground up. This complements regulation by demonstrating and institutionalizing best practices. For example, Canada's SCALE AI project funds both regulatory and capability-building efforts, providing shared infrastructure for data and training.

5.5. View 5: Public AI Will Be Inefficient and Capture-Prone.

There is a long history of inefficient or mismanaged public-sector technology projects. Bureaucracies move slowly, are vulnerable to capture, and can't attract talent (Mazzucato, 2013). Why should we expect public AI to be different?

Response: This is a valid concern. However, well-governed public institutions do exist and have produced extraordinary technological advances—from GPS to the internet to the Hubble Space Telescope. Moreover, public AI need not be synonymous with government-only models. Public funding could help support and scale up existing nonprofit activities—even OpenAI once contemplated asking for public funding (Klein, 2021). Proposals like Airbus for AI (Tan et al., 2025) envision a hybrid, multilateral structure of many national entities, each organized as public utilities. Successful examples like ERC, CERN, and W3C show that public AI can be designed to resist capture and reward quality. The risk of capture must be addressed, but the greater risk is failing to provide a meaningful alternative to concentrated private power.

6. Technical and Societal Implications

Public AI shifts the focus of machine learning research away from monolithic frontier labs and toward shared infrastructure, cooperative development, and inclusive deployment. For the ML community, this has far-reaching implications:

- For ML Researchers: Shared model libraries and pooled inference capacity democratize frontier experimentation. When models are shared, researchers can access and intervene on the internals of LLMs and other models without the cost or complexity of hosting their own hardware (Bommasani et al., 2021; Fiotto-Kaufman et al., 2025). They can access more of the RLHF and query data that is essential to frontier model capability research. Public AI also reduces fragility and promotes reproducibility across labs.
- For Non-CS Fields: Domains like healthcare, education, and law increasingly require high-quality models. Public AI enables domain experts to adapt systems to local needs without relying on private APIs or black-box deployments.
- For Open Source Ecosystems: Many contributors now work without guarantees that their outputs will remain in the commons. Public AI ensures their efforts resist private capture and support genuinely open systems.
- For Governments and Funders: Public AI can serve as a key plank of digital sovereignty and national innovation strategies (Public AI Network, 2024). Govern-

ments can focus investment on shared infrastructure and safety rather than competing on consumer UX.

- For the Broader Public: Public AI supports democratic accountability and contestability. It embeds collective input in how powerful systems are developed and used.

7. Conclusion

The machine learning community should not conflate open source with public good. We argue for a future in which open source AI is nested within public AI infrastructures: institutions and commitments that activate, sustain, and distribute AI systems for the public benefit.

This is not merely an ethical stance. It is a practical response to the material asymmetries of the current AI landscape. If the goal is to enable a diversity of actors to build and deploy capable models, then we must move beyond a romantic view of open source and begin investing in AI as public infrastructure.

If open source is to win, it must go public.

References

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M. M., Zhang, S., Ojewole, A., Guney, M. E., Biderman, S., Watkins, A. M., Ra, S., Lorenzo, P. R., Nivon, L., Weitzner, B., Ban, Y.-E. A., Chen, S., Zhang, M., Li, C., Song, S. L., He, Y., Sorger, P. K., Mostaque, E., Zhang, Z., Bonneau, R., and AlQuraishi, M. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8):1514–1524, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02272-z. URL https://www.nature.com/articles/s41592-024-02272-z. Publisher: Nature Publishing Group.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Wal, O. v. d. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, May 2023. URL http://arxiv.org/abs/2304.01373. arXiv:2304.01373 [cs].

BigScience Workshop, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., Moral, A. V. d., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., Strien, D. v., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., Toni, F. D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., ..., and Wolf, T. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, June 2023. URL http://arxiv.org/abs/2211.05100. arXiv:2211.05100 [cs].

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. GPT-NeoX-20B: An Open-Source Autoregressive Language Model, April 2022. URL http://arxiv.org/abs/2204.06745. arXiv:2204.06745 [cs].

Bommasani, R., Hudson, D. A., Adeli, E., and et al. On the opportunities and risks of foundation models. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2108.07258.

Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., and Liang, P. Considerations for governing open foundation models. *Science*, 386(6718):151–153, October 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adp1848. URL https://www.science.org/doi/10.1126/science.adp1848.

CAIRNE. Confederation of Laboratories for Artificial Intelligence Research in Europe. URL https://cairne.eu/cern-for-ai/.

Choksi, M. Z., Mandel, I., and Benthall, S. The brief and wondrous life of open models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3224–3240, 2025.

Collective Intelligence Project. A Roadmap to Democratic AI. Technical report, Collective Intelligence Project, March 2024. URL https://static1.squarespace.com/static/631d02b2dfa9482a32db47ec/t/65f9a1296f1a357e918f7a58/1722968408230/CIP_+A+Roadmap+to+Democratic+AI.pdf.

Eaves, D., Mazzucato, M., and Vasconcellos, B. Digital public infrastructure and public value: What is 'public'about dpi? 2024.

Eghbal, N. Roads and bridges: The unseen labor behind our digital infrastructure. Technical report, Ford Foundation, 2016. URL https://fordfoundation.org/m edia/2976/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure.pdf.

Ekgren, A., Gyllensten, A. C., Stollenwerk, F., Öhman, J., Isbister, T., Gogoulou, E., Carlsson, F., Heiman, A., Casademont, J., and Sahlgren, M. GPT-SW3: An Autoregressive Language Model for the Nordic Languages, October 2024. URL http://arxiv.org/abs/2305.12987. arXiv:2305.12987 [cs].

Fiotto-Kaufman, J., Loftus, A. R., Todd, E., Brinkmann, J., Pal, K., Troitskii, D., Ripa, M., Belfki, A., Rager, C., Juang, C., Mueller, A., Marks, S., Sharma, A. S., Lucchetti, F., Prakash, N., Brodley, C., Guha, A., Bell, J., Wallace, B. C., and Bau, D. NNsight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals, April 2025. URL http://arxiv.org/abs/2407.14561. arXiv:2407.14561 [cs].

Gerganov, G. ggml-org/llama.cpp, 2023. URL https://github.com/ggml-org/llama.cpp. original-date: 2023-03-10T18:58:00Z.

Gries, T. and Naudé, W. Modelling artificial intelligence in economics. *Journal for labour market research*, 56(1): 12, 2022.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. OLMo: Accelerating the Science of Language Models, June 2024. URL http://arxiv.org/abs/2402.00838. arXiv:2402.00838 [cs].

HAI, S. AI Index: State of AI in 13 Charts, April 2024. URL https://hai.stanford.edu/news/ai-index-state-ai-13-charts.

Hall, D., Ahmed, A., Chou, C., Garg, A., Kuditipudi, R., Held, W., Ravi, N., Shandilya, H., Wang, J., Bolton, J., Karamcheti, S., Kotha, S., Lee, T., Liu, N., Niklaus, J., Ramaswami, A., Salahi, K., Wen, K., Wong, C. H., Yang, S., Zhou, I., and Liang, P. Introducing marin: An open lab for building foundation models, May 2025. URL https://marin.community. Accessed: 2025-07-11.

Hugging Face. Model download statistics on hugging face hub (llama 3.1, olmo 2-7b, pythia). https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct,https://huggingface.co/allenai/OLMo-2-1124-7B,https://huggingface.co/EleutherAI/pythia-70m-deduped, 2025. Accessed May 21, 2025.

Johnson, I., Kaffee, L.-A., and Redi, M. Wikimedia data for ai: a review of wikimedia datasets for nlp tasks and ai-assisted editing. *arXiv preprint arXiv:2410.08918*, 2024.

Juijn, D., Pataki, B., Petropoulos, A., and Reddel, M. The EU's seat at the table. Technical report, Centre for Future Generations, September 2024. URL https://icfg.eu/wp-content/uploads/2024/09/CERN_for_AI_FINAL_REPORT.pdf.

Kelty, C. M. *Two Bits: The Cultural Significance of Free Software*. Duke University Press, June 2008. ISBN 978-0-8223-8900-2 978-0-8223-4242-7. doi: 10.2307/j.ctv1198vx9. URL http://www.jstor.org/stable/10.2307/j.ctv1198vx9.

Klein, E. Transcript: Ezra Klein Interviews Sam Altman. *The New York Times*, June 2021. ISSN 0362-4331. URL https://www.nytimes.com/2021/06/11/podcasts/transcript-ezra-klein-interviews-sam-altman.html.

Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T. Y., Wang, T., Dehaene, O., Davaadorj, M., Lamy-Poirier, J., Monteiro, J., Shliazhko, O., Gontier, N., Meade, N., Zebaze, A., Yee, M.-H., Umapathi, L. K., Zhu, J., Lipkin, B., Oblokulov, M., Wang, Z., Murthy, R., Stillerman, J., Patel, S. S., Abulkhanov, D., Zocca, M., Dey, M., Zhang, Z., Fahmy, N., Bhattacharyya, U., Yu, W., Singh, S., Luccioni, S., Villegas, P., Kunakov, M., Zhdanov, F., Romero, M., Lee, T., Timor, N., Ding, J., Schlesinger, C., Schoelkopf, H., Ebert, J., Dao, T., Mishra, M., Gu, A., Robinson, J., Anderson, C. J., Dolan-Gavitt, B., Contractor, D., Reddy, S., Fried, D., Bahdanau, D., Jernite, Y., Ferrandis, C. M., Hughes, S., Wolf, T., Guha, A., Werra, L. v., and Vries, H. d. StarCoder: may the source be with you!, December 2023. URL http://arxiv.org/abs/2305.06161. arXiv:2305.06161 [cs].

Luccioni, S., Jernite, Y., and Strubell, E. Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 85–99, 2024.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., Shi, A., and

Oak, S. Artificial intelligence index report 2025, 2025. URL https://arxiv.org/abs/2504.07139.

Mazzucato, M. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*. Anthem Press, 2013. ISBN 9780857282521.

Ollama Team. Ollama, May 2025. URL https://github.com/ollama/ollama. original-date: 2023-06-26T19:39:32Z.

OSI Opinion and Maris, J. Meta's LLaMa license is still not Open Source, February 2025. URL https://opensource.org/blog/metas-llama-license-is-still-not-open-source.

Paszke, A., Gross, S., Massa, F., and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019. URL https://arxiv.org/abs/1912.01703.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K. K., He, X., Hou, H., Lin, J., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., Mantri, K. S. I., Mom, F., Saito, A., Song, G., Tang, X., Wang, B., Wind, J. S., Wozniak, S., Zhang, R., Zhang, Z., Zhao, Q., Zhou, P., Zhou, Q., Zhu, J., and Zhu, R.-J. RWKV: Reinventing RNNs for the Transformer Era, December 2023. URL http://arxiv.org/abs/2305.13048. arXiv:2305.13048 [cs].

Public AI Network. Public AI: Infrastructure for the Common Good. Technical report, August 2024. URL https://zenodo.org/doi/10.5281/zenodo.13914560.

Raymond, E. S. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly Media, Sebastopol, CA, 1999. ISBN 9780596001087.

Reiss, J. Public Goods. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5B: An open large-scale dataset for training next generation image-text models, October 2022. URL http://arxiv.org/abs/2210.08402. arXiv:2210.08402 [cs].

Sitaraman, G. and Parek, K. The Global Rise of Public AI. Technical report, Vanderbilt Policy Accelerator, May 2025. URL https://cdn.vanderbilt.edu/vu-URL/wp-content/uploads/sites/412/2025/05/05220054/The-Global-Rise-of-Public-AI.pdf.

Tan, J., Jackson, B., Berjon, R., and Coyle, D. Airbus for AI: A Global Strategy for Public Value Creation. 2025.

Valero, M. and Crespo, D. Es la hora de 'AIbus': por qué Europa debe crear una gran empresa de AI. *El País*, April 2024. URL https://elpais.com/tecnologia/2024-04-11/es-la-hora-de-aibus-por-que-europa-debe-crear-una-gran-empresa-de-ai.html.

von Platen, P., Patil, S., Lozhkov, A., and et al. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Weber, M., Fu, D., Anthony, Q., Oren, Y., Adams, S., Alexandrov, A., Lyu, X., Nguyen, H., Yao, X., Adams, V., Athiwaratkun, B., Chalamala, R., Chen, K., Ryabinin, M., Dao, T., Liang, P., Ré, C., Rish, I., and Zhang, C. RedPajama: an Open Dataset for Training Large Language Models, November 2024. URL http://arxiv.org/abs/2411.12372. arXiv:2411.12372 [cs] version: 1.

Weber, S. *The Success of Open Source*. Harvard University Press, October 2005. ISBN 978-0-674-04499-9 978-0-674-01292-9. doi: 10.2307/j.ctv26071g2. URL http://www.jstor.org/stable/10.2307/j.ctv26071g2.

Widder, D. G., Whittaker, M., and West, S. M. Why 'open' AI systems are actually closed, and why this matters. *Nature*, 635(8040):827–833, November 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08141-1. URL https://www.nature.com/articles/s41586-024-08141-1. Publisher: Nature Publishing Group.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A. Comparing select open source software and open models

Properties	Linux	scikit-learn	TensorFlow	Kubernetes	OLMo	DeepSeek	LLaMA
Type	Operating System	ML Library	ML Framework	Container Orchestration	AI Model	AI Model	AI Model
Transparency	High — development is fully visible	High — all algorithms and tests are publicly documented and peer-reviewed	Medium — public codebase, but production usage depends on internal forks	High — development processes, governance, and roadmap are fully public	High — training pipeline, data decisions, and documentation openly shared	Medium — some training details and weights released, but pipeline unclear	Low — no access to training data, limited documentation, opaque post-training
Community Governance	Yes — community + Linux Foundation	Yes — consensus driven; backed by research orgs	Yes — SIGs, GitHub issues, TF RFCs	Yes — CNCF technical governance	Yes — AI2 hosts calls, roadmap, accepts contributions	No — releases set by DeepSeek	No — decisions made by Meta; no public forum
License Stability	Clear	Clear	Clear	Clear	Clear	Unclear	Unclear
Use Without Large Infra	Yes; runs on typical personal hardware	Yes; any Python env	Yes; CPUs or GPUs; many hosted options	Yes; single-node or small clusters	Partially; some GPUs locally; pruning supported	No; inference targets powerful clusters	No; needs high-end GPUs
Open Source Maintenance	Active; broad community + LF	Active; INRIA-led core + community	Active; Google + community	Active; CNCF + industry	Active; AI2 with public roadmap	Partial; periodic checkpoints	Irregular; Meta-driven
Business Model	Service-based (Red Hat, etc.), donations	Academic; grants/volunteers	Freemium support by Google	Cloud-vendor support via CNCF	Non-profit; philanthropic	Hedge-fund backed; opaque	Meta strategic positioning
Supporters/Adopters	Universities, enterprises, hobbyists, clouds	Universities, educators, research	Enterprises, researchers, hobbyists	Global enterprises, cloud providers	Academic labs, open-science advocates	Emerging China-centric dev community	Academic labs, startups via HF

Table 1. Comparing open source software and open source AI projects along a variety of axes, including transparency, governance, licensing, and maintenance concerns. A key takeaway is that AI is not like other open source software.