Distributed and Decentralised Training: Technical Governance Challenges in a Shifting AI Landscape

Jakub Kryś [1] **Yashvardhan Sharma** [1 2] **Janet Egan** [3]

Abstract

Advances in low-communication training algorithms are enabling a shift from centralised model training to compute setups that are either distributed across multiple clusters or decentralised via community-driven contributions. This paper distinguishes these two scenarios – distributed and decentralised training – which are little understood and often conflated in policy discourse. We discuss how they could impact technical AI governance through an increased risk of compute structuring, capability proliferation, and the erosion of detectability and shutdownability. While these trends foreshadow a possible new paradigm that could challenge key assumptions of compute governance, we emphasise that certain policy levers, like export controls, remain relevant. We also acknowledge potential benefits of decentralised AI, including privacy-preserving training runs that could unlock access to more data, and mitigating harmful power concentration. Our goal is to support more precise policymaking around compute, capability proliferation, and decentralised AI development.

1. Introduction

Large Language Model (LLM) training has rapidly grown in scale and now necessitates the use of massive computational resources (Sevilla & Roldán, 2024). Historically, this compute has been delivered in a single large data centre housing tightly interconnected GPUs. Such centralised settings are an attractive lever for AI governance as they are easily detectable and concentrated in the hands of a few entities (Sastry et al., 2024).

However, recent algorithmic developments indicate that it might be possible to train large-scale AI models on pools of compute that are not located in the same place and that can leverage slower connections. By drastically reducing the inter-GPU communication needs, these developments open the door to two new training paradigms: (1) hyperscalers performing training runs across multiple geographically dispersed clusters and (2) community-driven training of models over the Internet with no centralised point of control. Progress is happening fast on both fronts. Several frontier models have already been trained by combining compute from multiple data centres (Gemini Team et al., 2024; OpenAI, 2025). At the same time, since around the beginning of 2024 there has been a rapid rise of startups that use decentralised approaches to training AI models. We estimate these startups have raised around $145 million in venture capital funding so far[1], with the goal of training o3-level models by the end of 2025 (Prime Intellect, 2025b).

Both scenarios have important implications for the future of technical AI governance and we believe they should occupy more space in the policymaking discourse. As a first step, we offer the following contributions:

- We propose that the AI policy community clearly distinguish between the terms 'distributed' and 'decentralised' training to help delineate these scenarios and enable more precise discussion.
- We briefly describe the algorithmic progress enabling these developments, as well as their motivations and likely future trajectory.
- We detail several implications of this emerging trend on AI governance, including an increased risk of compute structuring and the erosion of detectability and shutdownability.
- We highlight possible benefits of decentralised training, including access to more data in a privacy-preserving manner and mitigation of power concentration risks.

[1]ML Alignment & Theory Scholar [2]Minerva University [3]Center for a New American Security. Correspondence to: Jakub Kryś <jakkrys@gmail.com>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

[1]This number is the aggregate funding received by the most prominent startups operating in this space: Nous Research – $50M (Trading View, 2025), Gensyn AI – $43M (Cryptopolitan, 2023), Prime Intellect – $20M (Prime Intellect, 2025e), Flower Labs – $20M (Labs, 2024), Pluralis Research – $7.6M (Crypto, 2025) and Berkeley Compute – $5M (Berkeley Compute, 2024).

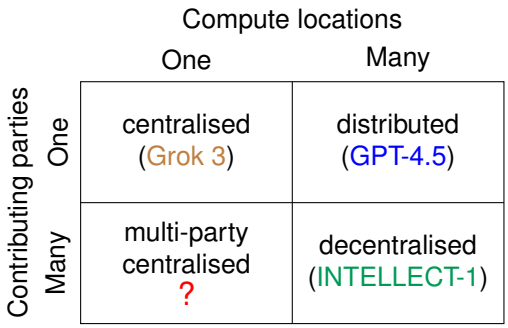

Figure 1. Possible training regimes organised by the number of contributing parties and compute locations. Each quadrant contains example models: Grok 3 represents physically centralised compute with a single coordinating party (xAI, 2025); GPT-4.5 illustrates centralised coordination across many compute sites (OpenAI, 2025); INTELLECT-1 exemplifies decentralised training with many contributors and locations (Jaghouar et al., 2024); and ? marks the hypothetical scenario of multiple parties contributing compute placed in one location.

2. Distributed and Decentralised Training

2.1. Distinction

We believe it is useful to clearly distinguish between two emerging paradigms: distributed and decentralised training (see Fig. 1). Historically, distributed training was a purely technical term referring to training models on several GPUs. Today, due to model size, every training run of frontier LLMs is distributed across many GPUs, rendering the original meaning obsolete. Thus, we propose the following terminology:

- **Distributed training** – training across multiple physically distant pools of compute, with a central entity coordinating the process. Other appropriate names could be 'multi-data centre training' or 'geographically distributed training'.
- **Decentralised training** – training that involves community-provided compute pools and contributions without a central coordinating entity.

Both paradigms leverage the same technical advances but differ in motivations and AI governance implications, warranting a clear distinction between them[2].

2.2. Distributed training

Building clusters large enough for frontier LLM training poses significant challenges, from permitting to accessing the energy needed to power them. Power demands for the largest training runs today are well above 100 megawatts – equivalent to supplying around 80,000 households with electricity – and are projected to grow to 5 gigawatts or above by 2030 (Sevilla et al., 2024; Fist & Datta, 2025; Pilz et al., 2025b). Given the difficulty of building and accessing energy at this scale in a single location, an alternative is to use multiple smaller data centres placed in different areas. Several AI labs already use multi-data centre training, for example in training GPT-4.5 and Gemini-1.5, although the exact distances between these data centres are not publicly known (Gemini Team et al., 2024; OpenAI, 2025). We anticipate that power constraints will continue to drive increased adoption of such distributed AI training.

To accomplish this, training runs can utilise so-called 'data parallelism', which is a method of training several copies of the model concurrently. Each copy receives a subset of the training data, updates its own parameters, then synchronises them with the other copies to produce an optimal model. To perform this synchronisation, GPUs have traditionally been interconnected through a very high bandwidth network, enabled by specialised equipment within a single cluster. Long-distance training is significantly more challenging, but analysis suggests that hyperscalers could achieve sufficient synchronisation by tapping into the vast amounts of unused fibre optic cables across the United States known as 'dark fibre' (Patel et al., 2024; Zayo, 2024). This could enable hyperscalers to train models across much greater distances.

2.2.1. Relevant algorithmic progress

New training methods that build on data parallelism can also help enable more ambitious uses of distributed training. Recently developed low-communication algorithms allow for much less intensive synchronisation of the data-parallel copies, often by as much as a factor of 500 (for a brief overview, see Appendix A) (Douillard et al., 2023; Liu et al., 2024; Douillard et al., 2025; Kale et al., 2025; Peng et al., 2024). In practical terms, this potentially enables conducting training runs over networks as slow as typical Internet speeds. This means that the importance of some types of interconnects could now be significantly lower, aiding efforts to synchronise GPUs less frequently and over longer distances. We will discuss the implications of this effect in Sec. 3.

2.3. Decentralised Training

This same algorithmic progress in low-communication training is also supporting new approaches to *decentralised* training. Several startups, most notably Prime Intellect and Nous Research, have already managed to pre-train models at the scale of 10 billion parameters on GPUs that were placed on different continents and communicated using typical In-

[2]At present, these terms are often used interchangeably (Bye, 2025), yet they pose distinct challenges. The case for making this distinction has also been made by (Lehman, 2025; KNOWER, 2025).

ternet speeds (Jaghouar et al., 2024; Nous Research, 2024). While these models currently underperform on benchmarks relative to similarly-sized counterparts, it is unclear to what extent this should be interpreted as a sign against their ultimate potential. For now, these efforts predominantly focused on demonstrating the feasibility of global decentralised training runs, rather than achieving optimal performance. Further improvements are likely achievable, though the ultimate competitiveness gap remains uncertain[3].

On the whole, the overarching goal of decentralised AI is to democratise access to AI, encourage open innovation and offer a counterbalance to the hyperscalers' dominance. The aim of these initiatives is to accumulate enough community-driven compute to enable the creation of collectively-owned, open-source models that can rival leading industry products. Indeed, several precedents for globally pooled compute already exist, for example the Folding@Home project which at its peak accumulated 280 000 GPUs and was the first compute pool to cross the 10^{18} FLOPS threshold (Zimmerman et al., 2021)[4]. However, it should be noted that AI training (and inference) involves very different computational workloads, so this raw FLOPS count should not be directly translated to the AI context. Nonetheless, it illustrates the rough magnitude of decentralised compute that could be accumulated. We leave a precise estimate of the largest possible decentralised training run to future work.

2.3.1. Peer-to-Peer Communication

Unlike the multi-data centre distributed training described in Sec. 2, decentralised training runs do not rely on a single point of control. Most GPU network layouts involve at least one type of centralisation: (1) algorithmic centralisation, where model synchronisation happens through a single server, or (2) operational centralisation, where an organisational layer coordinates what each GPU should do, monitors their status and manages resources. The opposite design can be implemented in so-called peer-to-peer (P2P) networks. Here, there is no central choke point and all tasks are initiated and coordinated by the network nodes themselves. When training an AI model, new nodes can contribute to the process *trustlessly*, that is without an approval by a central authority. The legitimacy of their work is attested through cryptographic proofs and verified by other nodes (Prime Intellect, 2025d). Similarly, model synchronisation can be done without a master node, instead relying on peers coordinating with other peers directly (Ryabinin et al., 2021). Even if a given GPU fails, this does not halt the training run – the network can dynamically reorganise, continue training, and accept replacement GPUs that obtain key instructions and information from their peers (Ryabinin et al., 2020; Jaghouar et al., 2024).

In terms of design, this setup is much more closely related to the decentralised structure of blockchains such as Ethereum than it is to traditional data-parallel training. While no training protocol has achieved a *fully* decentralised status as of today, this is a stated goal (Prime Intellect, 2025c). This could signal the emergence of a new regime, one where oversight is much harder and some of the key assumptions underpinning compute governance are questioned.

3. Implications for Governance

In this section, we discuss the implications of these trends for AI governance. Advances in distributed training will likely enable leading AI companies to bypass energy constraints and continue to scale, reinforcing their dominance at the forefront of AI. Decentralised training, however, comes with a novel set of challenges and opportunities.

3.1. Capability Proliferation

Decentralised AI – alongside distillation, algorithmic progress and hardware efficiency – could be one of the major drivers of AI capability proliferation (Pilz et al., 2025a). This trend could be especially significant if traditional open-weight providers shift to closed-weight deployment, curtailing the usual routes for unrestricted access. Moreover, it offers an alternative path for a broader set of actors to contribute their compute (potentially including consumer GPUs) to develop and deploy advanced AI systems. Notably, the post-training of models to unlock stronger capabilities is particularly well-suited to distributed and decentralised environments, which favour techniques that operate effectively under limited communication (see Appendix B). We therefore expect that access to, and ownership of, increasingly powerful models will become more commoditised. While it is unclear to what extent this trend will scale, progress since the beginning of 2024 has been impressive and suggests this issue deserves further attention of the AI governance community. We echo calls from other AI researchers to prioritise research on societal resilience (Toner, 2025).

Technical feasibility aside, the motivation to make decentralised AI happen is apparent – we should not be surprised if the next few years lead to a real explosion in uptake of these technologies, similarly to blockchain-style networks a decade ago, and with no clear way of overseeing or regulat-

[3] In particular, recent benchmarking results of reasoning models have questioned the quality of INTELLECT-2, a reasoning model post-trained in a decentralised manner, which seems to underperform relative to its base model (Hochlehnert et al., 2025). However, it is unclear whether this effect is due to its decentralised training procedure or other methodological flaws that apply to RL post-training more broadly.

[4] The project used around 280 000 GPUs and 4.8 million CPUs at its peak, but after accounting for their relative performance, we deduce that the GPUs were responsible for around 94% of the total FLOPS.

ing them. We do not take a stance on whether decentralising AI development is positive or not, but we call for continued evaluation of its risk-benefit profile through the 'marginal risk' framework (Kapoor et al., 2024). If decentralised AI significantly lags behind the frontier and sufficient defensive technologies can be developed, government intervention may be unnecessary, especially given the potential benefits of decentralisation.

3.2. Shutdownability

Decentralised training could undermine policy interventions that aim to promote 'shutdownability' of powerful AI models. To manage potentially catastrophic AI risks, it has been argued that we should preserve an option of shutting down training and inference of AI models, in case they are misaligned or are being misused (Barnett & Scher, 2025; United Nations, 2023). While open-weight models do make this more difficult, their release is not equivalent to a large-scale diffusion of capabilities. This is because in order to run them at scale, one still needs access to specialised compute that is typically housed in centralised clusters and costs more than what most actors can afford. This compute access provides a concrete point of intervention for AI governance. In contrast, by creating a true P2P network as described in Sec. 2.3.1, it might be possible to train or run inference on models in a fully decentralised manner and with no obvious point of accountability or intervention. This problem is further exacerbated in the context of agentic systems (Thornley, 2024). While there are important differences between decentralised training and inference, work on the latter is also actively underway (Prime Intellect, 2025a). The 'no off-switch' problem has been recognised within the decentralised AI community, albeit without any clear solution for now (Long, 2024).

3.3. Compute Structuring

The methods underpinning both distributed and decentralised training could increase the risk of compute structuring as way of avoiding government oversight. Compute thresholds based on the amount of pre-training FLOPS have been proposed as part of several regulatory frameworks (European Parliament, 2024). They rely on the assumption that since compute amounts serve as a rough proxy for the capabilities of the model, they can also indirectly serve as a way of targeting policy interventions towards the riskiest models. To evade such interventions, a malicious actor could divide their training run into several smaller workloads that are executed in parallel using different cloud providers. Such 'compute structuring' utilises data parallelism and so greatly benefits from low-communication methods that reduce synchronisation frequency. Detecting parallel structuring might be possible by monitoring the data volume and IP addresses that a compute pool communicates with, but will be challenging without information sharing between various cloud providers (Egan & Heim, 2023).

Unfortunately, work on KYC schemes and workload monitoring is not progressing at an adequate pace to mitigate this risk. Moreover, strategies to mask communication patterns are likely possible. For example, during the training of INTELLECT-1 (Jaghouar et al., 2024) over the Internet, synchronisation occurred only once every 38 minutes, and it is possible to purposefully prolong it by up to several training steps, thereby obfuscating any characteristic patterns. Overall, the advent of low-communication training algorithms makes parallel structuring significantly more feasible than before, even if we are uncertain as to how likely it is to be attempted in practice. For concrete proposals on preventing it, see (Seferis & Fist, 2025).

3.4. Relevance of Compute

Our analysis should not be interpreted as suggesting that the importance of compute quantity and quality has been rendered obsolete. Low-communication training changes the way compute can be leveraged, but does not imply that less of it is needed. While more actors may now be able to perform advanced AI training, front-runners in this field will still benefit from vast numbers of GPUs.

Even after accounting for decentralised training, it is likely that hyperscalers will continue to lead in developing and deploying the most advanced AI models. Any innovation that enhances decentralised training can typically be swiftly adopted by these actors if it offers advantages over their existing methods. In addition, hyperscalers might be able to preserve their competitive moat due to other constraints that continue to bottleneck training and advantage actors with deeper pockets. For instance, ultra-fast GPU memory – often accounting for nearly half the cost of the most advanced chips – can make such hardware prohibitively expensive for most (Prickett Morgan, 2024). Dominance in compute, whether at a company or country level, will continue to offer significant benefits to AI leadership. As such, we do not expect the relevance of compute (or interventions such as export controls on compute) to be diminished, at least in the foreseeable future.

3.5. Benefits of Decentralisation

Democratisation of access to AI systems could be a powerful way of avoiding economic and political disempowerment of individuals and nations (Kulveit et al., 2025). Some argue that in order to achieve this goal, we need a widespread diffusion of capabilities and cannot rely on centralised AI providers (Laine & Drago, 2025). Cheap, collectively-owned models could provide a solution.

Decentralisation could also significantly advance AI

progress by unlocking access to privately held data. While current models primarily utilise public information, vastly larger amounts of private data exist – such as conversations, car telemetry or CCTV recordings – that owners are hesitant to share with traditional AI developers. However, privacy-preserving ways of training already exist, and could be supported by decentralised approaches to leverage data without ever transferring it to a central entity. This could lead to models that are more personalised, better reflect local contexts and are tailored to novel applications (Sani et al., 2024; Vana, 2025). This is an area where decentralised training may have a comparative advantage over traditional approaches. Moreover, since data does not need to cross jurisdictions, this could also support international cooperation on AI efforts while maintaining data sovereignty.

4. Conclusions

In this paper, we have presented recent developments in the domain of low-communication training algorithms and described two new paradigms that emerge as a result. We argued for making the distinction between multi-data centre *distributed* training and *decentralised* training, which will enable more precise discourse. We discussed several implications of these developments, from possible capability proliferation to the 'no off-switch' problem, but also the potential creation of more accessible, tailored AIs.

There are several ways to take this work forward:

- Creating robust methods of detecting parallel compute structuring rises in importance. This could be done through a combination of KYC schemes, workload monitoring and on-chip mechanisms, but work in this field is preliminary and should be accelerated (Seferis & Fist, 2025).
- It is unclear to what extent low-communication training can keep up with the frontier of traditional training performed by hyperscalers. Open questions remain about whether improvements similar to DiLoCo can be achieved in other types of parallelism, such as tensor or pipeline parallelisms. It would be especially useful to adapt existing simulation tools to include low-communication parallelism of all types, even if they remain speculative for now (Erdil & Besiroglu, 2024; Douillard et al., 2025; Exo Labs, 2025).
- Likewise, governance discourse would benefit from a precise estimate of computing power that could be pooled together from consumer GPUs, as well as data centre GPUs in small- to medium-sized clusters. This would enable projecting the scale of decentralised training runs more precisely.
- In terms of regulatory interventions such as export controls, it may be useful to include a focus on intra-node networking equipment, as well as GPU memory bandwidth and latency. These could offer more precise control over the use of chips for training at scale.

Impact Statement

This paper presents technical insights into Machine Learning to enable better AI governance and policymaking. We hope it contributes to a balanced discussion of the risks and benefits of powerful AI systems.

Acknowledgments

We would like to thank Matthew Wearden, Kevin Wei, Nathan Helm-Burger, Zilan Qian, as well as anonymous ICML reviewers for helpful comments and discussions. JK graciously acknowledges the financial support of Open Philanthropy.

References

Barnett, P. and Scher, A. AI governance to avoid extinction: The strategic landscape and actionable research questions, 2025. URL https://techgov.intelligence.org/research/ai-governance-to-avoid-extinction.

Berkeley Compute. Introducing Berkeley Compute: Revolutionizing AI with decentralized GPU power, 2024. URL https://blog.berkeleycompute.com/introducing-berkeley-compute/.

Bye, L. Decentralized training isn't a policy nightmare – yet, 2025. URL https://www.transformernews.ai/p/decentralized-training-policy-implications.

Charles, Z., Teston, G., Dery, L., Rush, K., Fallen, N., Garrett, Z., Szlam, A., and Douillard, A. Communication-efficient language model training scales reliably and robustly: Scaling laws for DiLoCo. *arXiv preprint arXiv:2503.09799*, 2025.

Crypto, F. Exclusive: Pluralis raises $7.6 million from prominent investors to take on OpenAI with big decentralized models, 2025. URL https://fortune.com/crypto/2025/03/19/pluralis-research-7-6-million-openai-coinfund-union-square-ventures/.

Cryptopolitan. Gensyn AI raises $43 million in Series A funding led by a16z to revolutionize decentralized machine learning, 2023. URL https://www.cryptopolitan.com/gensyn-ai-raises-43-million-in-series-a-funding-led-by-a16z/.

Douillard, A., Feng, Q., Rusu, A. A., Chhaparia, R., Donchev, Y., Kuncoro, A., Ranzato, M., Szlam, A., and Shen, J. DiLoCo: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023.

Douillard, A., Donchev, Y., Rush, K., Kale, S., Charles, Z., Garrett, Z., Teston, G., Lacey, D., McIlroy, R., Shen, J., et al. Streaming DiLoCo with overlapping communication: Towards a distributed free lunch. *arXiv preprint arXiv:2501.18512*, 2025.

Egan, J. and Heim, L. Oversight for frontier AI through a know-your-customer scheme for compute providers. *arXiv preprint arXiv:2310.13625*, 2023.

Erdil, E. and Besiroglu, T. Introducing the distributed training interactive simulator, 2024. URL https://epoch.ai/blog/introducing-the-distributed-training-interactive-simulator.

European Parliament. EU AI Act: Article 51 – classification of general-purpose AI models as general-purpose AI models with systemic risk, 2024. URL https://artificialintelligenceact.eu/article/51/.

Exo Labs. Exo Gym: Open-source Python toolkit to facilitate distributed AI research., 2025. URL https://github.com/exo-explore/gym. Accessed: 2025-05-02.

Fist, T. and Datta, A. Compute in America: A policy playbook, 2025. URL https://ifp.org/special-compute-zones/.

Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., and Bethge, M. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

Jaghouar, S., Ong, J. M., Basra, M., Obeid, F., Straube, J., Keiblinger, M., Bakouch, E., Atkins, L., Panahi, M., Goddard, C., et al. INTELLECT-1 technical report. *arXiv preprint arXiv:2412.01152*, 2024.

Kale, S., Douillard, A., and Donchev, Y. Eager updates for overlapped communication and computation in DiLoCo. *arXiv preprint arXiv:2502.12996*, 2025.

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., et al. On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*, 2024.

KNOWER. Unpacking decentralized training, 2025. URL https://theknower.substack.com/p/unpacking-decentralized-training.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental AI development. *arXiv preprint arXiv:2501.16946*, 2025.

Labs, F. Flower Labs raises $20M Series A, 2024. URL https://flower.ai/blog/2024-02-15-announcing-series-a/.

Laine, R. and Drago, L. The intelligence curse, 2025. URL https://intelligence-curse.ai/breaking/.

Lehman, S. Unpacking decentralized training, 2025. URL https://www.symbolic.capital/writing/frontier-training.

Liu, B., Chhaparia, R., Douillard, A., Kale, S., Rusu, A. A., Shen, J., Szlam, A., and Ranzato, M. Asynchronous local-sgd training for language modeling. *arXiv preprint arXiv:2401.09135*, 2024.

Long, A. Protocol learning, decentralized frontier risk and the no-off problem. *arXiv preprint arXiv:2412.07890*, 2024.

Nous Research. Nous DisTrO, 2024. URL https://distro.nousresearch.com/.

OpenAI. OpenAI GPT-4.5 System Card, 2025. URL https://openai.com/index/gpt-4-5-system-card/.

OpenAI. Introduction to GPT-4.5. https://www.youtube.com/live/cfRYp0nItZ8?t=457s, 2025. Accessed April 29, 2025. Quoted at 7:37 (457 seconds).

Patel, D., Nishball, D., and Ontiveros, J. E. Multi-datacenter training: OpenAI's ambitious plan to beat Google's infrastructure, 2024. URL https://semianalysis.com/2024/09/04/multi-datacenter-training-openais/.

Peng, B., Quesnelle, J., and Kingma, D. P. Decoupled momentum optimization. *arXiv preprint arXiv:2411.19870*, 2024.

Pilz, K. F., Heim, L., and Brown, N. Increased compute efficiency and the diffusion of AI capabilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27582–27590, 2025a.

Pilz, K. F., Sanders, J., Rahman, R., and Heim, L. Trends in AI supercomputers, 2025b. URL https://arxiv.org/abs/2504.16026.

Prickett Morgan, T. He who can pay top dollar for HBM memory controls ai training, 2024. URL https://www.nextplatform.com/2024/02/27/he-who-can-pay-top-dollar-for-hbm-memory-controls-ai-training/.

Prime Intellect. Planetary-scale inference: Previewing our peer-to-peer decentralized inference stack, 2025a. URL https://www.primeintellect.ai/blog/inference.

Prime Intellect. Decentralized training in the inference-time-compute paradigm, 2025b. URL https://www.primeintellect.ai/blog/intellect-math.

Prime Intellect. Introducing Prime Intellect's protocol & testnet: A peer-to-peer compute and intelligence network, 2025c. URL https://www.primeintellect.ai/blog/protocol.

Prime Intellect. Toploc - a locality sensitive hashing scheme for trustless verifiable inference, 2025d. URL https://www.primeintellect.ai/blog/toploc.

Prime Intellect. $15m to build a peer-to-peer AI protocol, 2025e. URL https://www.primeintellect.ai/blog/fundraise.

Ryabinin, M., Borzunov, A., Diskin, M., Gusev, A., Mazur, D., Plokhotnyuk, V., Bukhtiyarov, A., Samygin, P., Sinitsin, A., and Chumachenko, A. Hivemind: Decentralized Deep Learning in PyTorch, April 2020. URL https://github.com/learning-at-home/hivemind.

Ryabinin, M., Gorbunov, E., Plokhotnyuk, V., and Pekhimenko, G. Moshpit SGD: Communication-efficient decentralized training on heterogeneous unreliable devices. *Advances in Neural Information Processing Systems*, 34: 18195–18211, 2021.

Sani, L., Iacob, A., Cao, Z., Marino, B., Gao, Y., Paulik, T., Zhao, W., Shen, W. F., Aleksandrov, P., Qiu, X., et al. The future of large language model pre-training is federated. *arXiv preprint arXiv:2405.10853*, 2024.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., et al. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*, 2024.

Seferis, M. and Fist, T. Detecting compute structuring in AI governance is likely feasible, 2025.

Sevilla, J. and Roldán, E. Training compute of frontier AI models grows by 4-5x per year, 2024. URL https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year. Accessed: 2025-04-29.

Sevilla, J., Besiroglu, T., Cottier, B., You, J., Roldán, E., Villalobos, P., and Erdil, E. Can AI scaling continue through 2030?, 2024. URL https://epoch.ai/blog/can-ai-scaling-continue-through-2030. Accessed: 2025-04-29.

Thornley, E. The shutdown problem: an AI engineering puzzle for decision theorists. *Philosophical Studies*, pp. 1–28, 2024.

Toner, H. Nonproliferation is the wrong approach to AI misuse, 2025. URL https://helentoner.substack.com/p/nonproliferation-is-the-wrong-approach.

Trading View. Nous Research secures $50m from Paradigm to build decentralized AI on Solana, 2025. URL https://www.tradingview.com/news/cointelegraph:54f1a44d2094b:0-nous-research-secures-50m-from-paradigm-to-build-decentralized-ai-on-solana/.

United Nations. Interim report: Governing AI for humanity, 2023. URL https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf.

Vana. Vana and Flower Labs partner to build the world's first user-owned AI model, 2025. URL https://www.vana.org/posts/vana-flower-labs-partnership.

xAI. Grok 3 Beta — the age of reasoning agents, 2025. URL https://x.ai/news/grok-3.

Zayo. A guide to dark fiber networks, 2024. URL https://www.zayo.com/resources/a-guide-to-dark-fiber/.

Zimmerman, M. I., Porter, J. R., Ward, M. D., Singh, S., Vithani, N., Meller, A., Mallimadugula, U. L., Kuhn, C. E., Borowsky, J. H., Wiewiora, R. P., Hurley, M. F. D., Harbison, A. M., Fogarty, C. A., Coffland, J. E., Fadda, E., Voelz, V. A., Chodera, J. D., and Bowman, G. R. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chemistry*, 13(7):651–659, July 2021. ISSN 1755-4349. doi: 10.1038/s41557-021-00707-0. URL https://doi.org/10.1038/s41557-021-00707-0.

A. Low-Communication Data-Parallel Training

Data Parallelism

In the context of LLM training, data parallelism refers to the strategy of distributing a model's training workload across multiple processing units. Each unit – typically a GPU or a group of GPUs – holds a full copy of the model and processes a distinct 'shard' (a subpart) of the dataset. After computing local gradient updates based on its assigned shard, each unit shares these gradients (or the updated weights) with the others. The updates are then averaged to maintain consistency across the model replicas (see Fig. 2).

Traditionally, this synchronisation step occurs after every training batch, which guarantees strict consistency, but imposes significant communication costs. For example, if processing a batch of tokens takes 4 seconds, while synchronisation requires an additional second, then 20% of total training time is lost to communication. In large-scale settings, where models might occupy hundreds of gigabytes, this becomes a bottleneck even within a single data centre – and entirely prohibitive across multiple data centres.

Challenges of High-Frequency, High-Bandwidth Synchronisation

Standard data parallelism assumes fast, stable interconnects like NVLink or InfiniBand. Extending it to multi-cluster settings quickly becomes unviable due to synchronisation frequency and peak bandwidth requirements. Even if the connection is idle most of the time, it must support extremely high peak throughput when communication occurs. To illustrate the problem, suppose model synchronisation takes 60 seconds over the public Internet (30 seconds both ways), and local training still takes 4 seconds per batch. If synchronisation occurs after every batch, compute utilisation drops to just 6.25%. At this point, most of the GPU's time is spent waiting for data rather than training the model. Reducing the synchronisation frequency alleviates this problem, but naive approaches severely degrade training performance due to diverging model replicas and outdated gradients.

Emergence of Low-Communication Training Methods

Recent algorithmic work – most notably DiLoCo (Distributed Low Communication), DeMo (Decoupled Momentum), and their successors – represents a significant breakthrough (Douillard et al., 2023; Liu et al., 2024; Douillard et al., 2025; Kale et al., 2025; Peng et al., 2024). These methods either modify the training protocol to allow workers to synchronise infrequently (as little as once every 500 steps), or with the same frequency, yet transferring a much smaller volume of data.

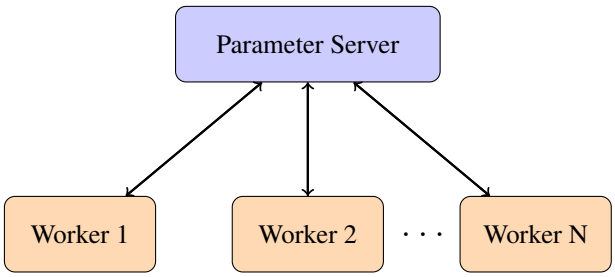

Figure 2. In data-parallel training, there are N workers which train local replicas of the model on their assigned data shards. After updating the weights, these changes need to be synchronised across all workers. In the most basic implementation, this can be done through a single 'parameter server', but more efficient strategies are used in practice.

For example, the core idea of DiLoCo is to detach local optimisation from global synchronisation through the use of two separate optimisers. One optimiser (often a variant of stochastic gradient descent) handles the local model updates, while a second one adjusts the global parameters after periodic synchronisation. This design maintains stability and convergence even when dealing with outdated gradients. Early empirical studies indicate that such low-communication methods achieve comparable or even better performance than their traditional counterparts.

This training protocol leads to very large gains in compute utilisation. Using the previous example of training over the Internet, synchronising model replicas every 500 steps (∼33 minutes) instead of every step improves compute utilisation from 6.25% to over 97%. In practical terms, this enables meaningful training runs over typical Internet connections – a key requirement of decentralised AI.

Additional Benefits and Follow-Up Improvements

Apart from reducing communication needs, DiLoCo-style algorithms have other important benefits. Empirical work has demonstrated several properties that make them very well-suited for decentralised and multi-data centre training:

- **Asynchronous communication**: DiLoCo-style optimisers are robust to asynchronous averaging, that is situations where only a subset of replicas can average their parameters at a given point. This property is useful from the perspective of training over unstable connections, where individual contributors may drop in and out of the compute pool.
- **Robustness to data shards from different distributions**: performance does not degrade when workers train on datasets that follow distinct underlying distributions (for instance, with varying numbers of training samples related to a certain feature). This is key to

unlocking vast amounts of privately-held data, which naturally will have a high degree of variation between individual contributors.

- **Improved performance with model scale**: perhaps most importantly, follow-up work by (Charles et al., 2025) showed that in certain settings, DiLoCo and its successors can *outperform* traditional data-parallel training altogether – with improvements in both convergence speed and final loss. Their analysis also found that the effectiveness of these methods improves with model size, suggesting that low-communication training might be able to keep up with the scaling paradigm.

Overall, these properties are in stark contrast to normal data parallelism, which is typically fragile to node failures, requires homogeneous hardware, and depends on continuous high-speed connectivity.

B. Reasoning Models and Decentralised Compute

Since late 2024, there has been an emergence of so-called reasoning models, such as OpenAI's 'o' series, Claude 3.7 Sonnet-thinking, and DeepSeek-R1. These models are deliberately taught to 'think' before producing the final answer, typically generating hundreds or thousands of tokens during an interaction. This implies a rebalancing in AI workloads, since serving the model to customers now requires a growing amount of compute – a phenomenon often referred to as 'inference-time scaling'. Crucially, it also bears important implications for their training process. Below, we explain why reasoning models can more easily accommodate low-bandwidth environments during post-training than traditional models, making them more compatible with distributed and decentralised setups.

Reasoning Models Are Suited for Low Communication Environments

Most reasoning models are post-trained using reinforcement learning (RL), which differs in several fundamental ways from the self-supervised learning (SSL) used in pre-training. SSL involves generating predictions for a known sequence of tokens (known as a 'forward pass'), comparing those predictions to the ground truth, and adjusting the model weights such that the next prediction is incrementally better ('backward pass'). Importantly, this process involves a 1:1 ratio of forward and backward passes – every forward pass must be accompanied by a weight update. As discussed in Appendix A, synchronising these updates across all nodes in a network is highly communication-intensive and generally requires fast, stable connectivity.

In contrast, post-training with RL involves a much larger number of forward passes per gradient update. The model explores a large space of possible 'thinking trajectories', each consisting of many new tokens generated one at a time. These trajectories are scored using a reward function, and only then used to perform an update to the model weights. In this setup, the ratio of forward to backward passes can easily reach 1000:1. This naturally reduces the frequency of weight updates, which in turn reduces the frequency of synchronisation. In settings with limited bandwidth, such as those encountered in distributed or decentralised training, this is a highly desirable property.

Practical Implications

The shift to reasoning models lowers the bar for contributing compute to advanced AI model training. Normally, in order to take part in a training run, we need to have enough compute to fit the model into memory, calculate the gradients in a backward pass, and synchronise them with other nodes in the network. However, generating thinking trajectories for RL-based post-training is less compute intensive and requires less memory, opening the door to the usage of consumer-grade hardware. To give a concrete example, a single M3 Ultra platform by Apple, which comes with up to 512 gigabytes of memory, is sufficient to run inference on models as large as DeepSeek-R1 (671 billion parameters). Moreover, such nodes do not need to be connected to their peers with high-bandwidth networks, as they only have to communicate the generated thinking trajectories and their corresponding rewards, not the full gradients. Thus, this aspect of RL training can be distributed (or even decentralised) across thousands of dispersed contributors, each of whom can utilise typical Internet connections.