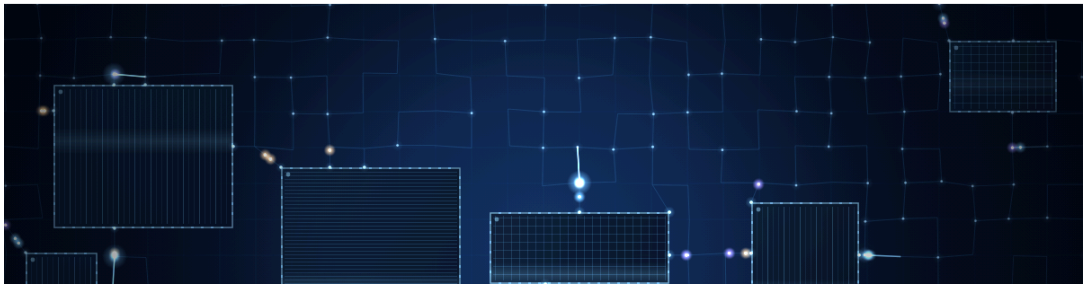
AI Hardware Choices are Highly Variable and Sparsely Disclosed



AUTHORS

Francisco Ríos *, Ian Reynolds *, Robert Praas , Avijit Ghosh †, Irene Solaiman †

* These authors contributed equally.

† Advisors.

AFFILIATIONS

Hugging Face

Centre for European Policy Studies

PUBLISHED

Jun. 04, 2026

Table of Contents

Executive Summary

Hardware concerns are critical for understanding global AI development trends, yet we have minimal insight into the hardware configurations enabling model training and inference. We analyzed the top 4,000 most downloaded models on Hugging Face and point to three main trends with respect to AI training and inference hardware:

1. There is a notable lack of transparency in hardware disclosure. On the training side, our analysis found that more than 40% give no recoverable public signal of the chips used to train them. The picture is just as opaque with respect to inference. This is important as model performance is tightly coupled to the hardware it runs on. As a result, a mismatch between training and inference hardware produces measurable behavioral divergence.
2. Model training choices are diversifying. Among developers that disclose training hardware NVIDIA still dominates across both US and Chinese labs. That said, we find Chinese developers increasingly mix in domestic accelerators, notably Huawei's Ascend series.
3. Inference options have expanded. This is particularly relevant in the context of locally deployable models, which are increasingly accessible and help users avoid high token costs and GPU constraints.

These findings have technical and policy consequences. For the technical community, it means benchmark numbers cannot be read in isolation, because a reported accuracy figure reflects the serving setup and ecosystem reverberations as much as the model itself. For policymakers, hardware choices and their disclosure better establishes the baselines needed to assess global hardware dependencies, supply chain exposure, and the state of international compute competition with sufficient empirical grounding. Both training and inference hardware should be treated as standard elements of model release documentation.

Introduction

Hardware concerns, particularly advanced chips, have become a critical focal point within the AI ecosystem, with serious signs pointing toward a compute crunch where demand outpaces supply. Access to this hardware determines the baseline resources available to developers and deployers, while also shaping the landscape of international technology competition. For example, US export controls on advanced chips are seen as a key tool for limiting Chinese AI advances. Chinese companies are attempting to onshore chip manufacturing to avoid supply

constraints, and are reportedly acquiring advanced controlled chips illicitly. Private companies are also trying to secure more access by developing their own in-house production efforts. Moreover, compute constraints continue for model training and inference, meaning additional chips are needed to fuel further model development and deployment.

Despite these overarching trends, we thus far have limited visibility into data on which models are trained on which hardware. This applies in both the open-source and closed-source ecosystems. Insight into this information is important; as laid out in more detail below, the relationship between a model's training hardware and its downstream inference performance can be highly variable.

Without establishing the baseline hardware environment a model was trained on, it is difficult to evaluate, diagnose, or optimize its behavior when deployed across different infrastructure. Mapping training hardware serves a few valuable purposes. Improved insight into hardware trends in training and inference can underscore evolving preferences within the AI community, better informing user and developer compute decisions, optimizations, investments, and trends in efficiency and cost preference. Moreover, these trends help to illustrate widening geopolitical divides and increasing competition driven by hardware concerns. As a result, this analysis can help the technical and policy communities better understand user preferences, guide infrastructure development trends, and make more informed policy decisions.

In light of this gap in data, this blog post makes three core points:

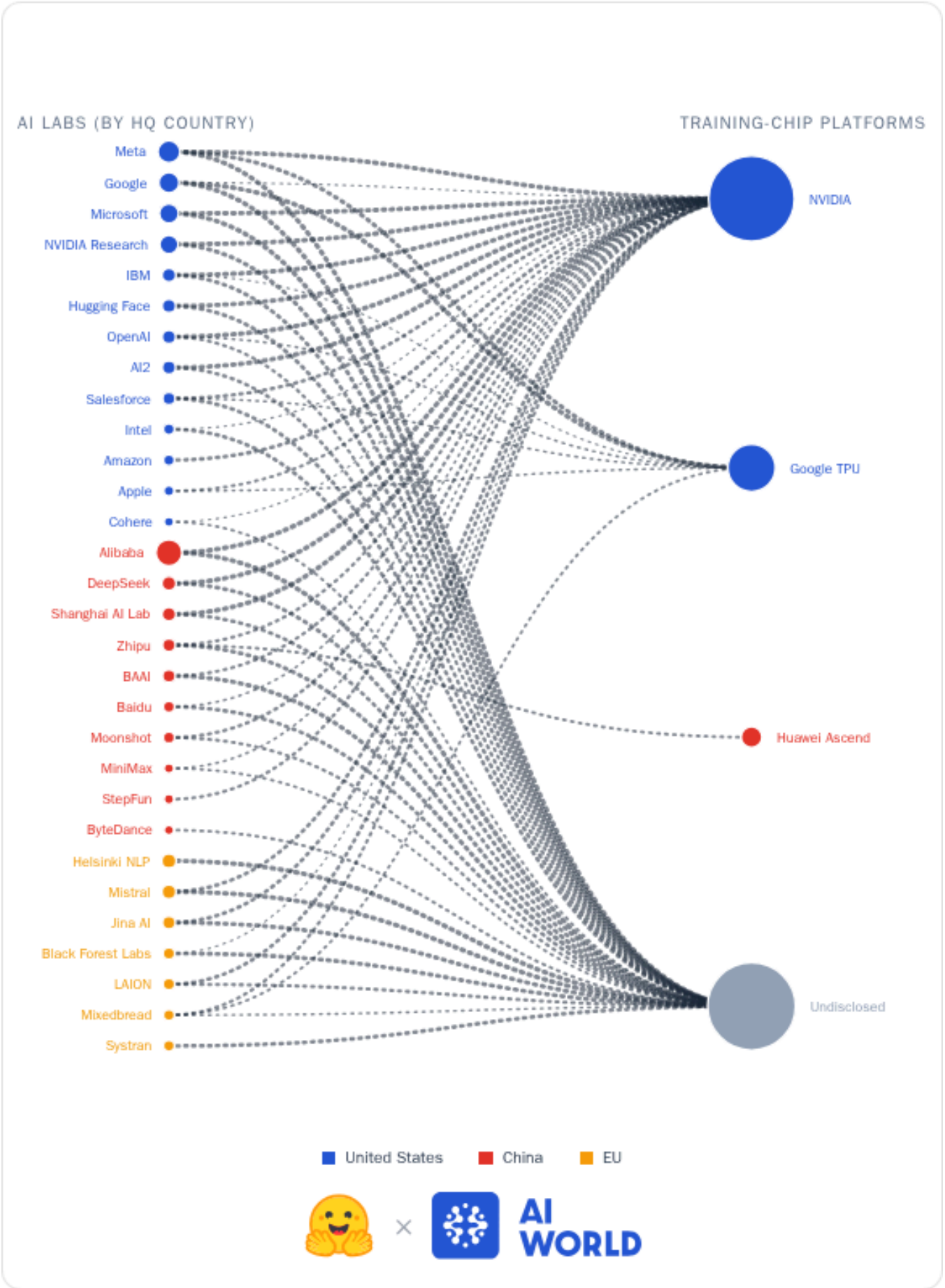
- Training: we map out the currently available data on which specific models are trained on which chips, while highlighting the general lack of transparency in this space across both the open- and closed-source ecosystems. Across our analysis, NVIDIA appears in the training stack of nearly every major lab, while over 40% of models give no recoverable signal of the hardware used.
- Inference: top-end machine learning performance across both US and Chinese chips is steadily improving, with US chips continuing to outperform their Chinese counterparts. Yet there appears to be a growing interest in diverse hardware for inference, especially for facilitating model deployment on Chinese chips, which has downstream performance implications due to hardware mismatch.
- Organizational: increased transparency, by making disclosure of hardware the norm, allows for better analysis of ecosystem trends and the global AI development dynamics that carry ongoing political implications.

Mapping the Model Training Hardware Environment

To our knowledge, this is the first systematic audit of training hardware disclosure across a large sample of open-weight models. We analyzed the hardware used for training the top 4,000 most-downloaded models on the Hugging Face platform as of early May 2026. The main data sources consulted were the Hugging Face model card, the GitHub page, and potential attached arXiv papers. For the most common US and Chinese chip providers we selected signals, such as Trainium for Amazon and CUDA for NVIDIA; a more detailed methodology is in the appendix.

Figure 1 shows that among top US and Chinese model developers, many have trained at least one model on NVIDIA chips, while TPUs are mostly used by Google among a few others. Among notable labs, Zhipu AI is the only one we could record using Huawei Ascend in disclosed training documentation. Other labs, such as DeepSeek, have been widely reported to train on Ascend hardware but have not disclosed this in their model cards or technical reports. Some of this information can be found only in Chinese-language announcements.

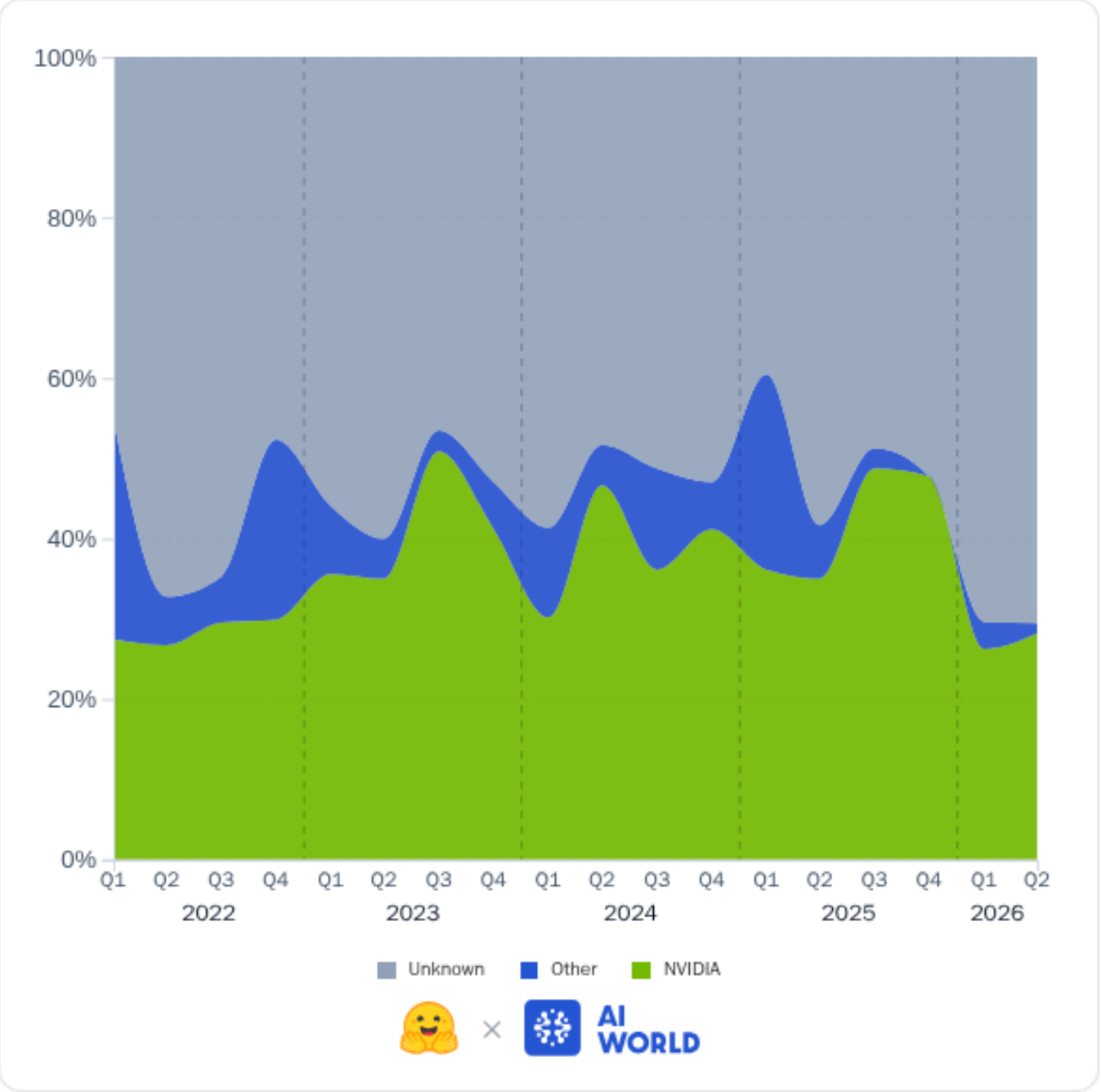
Figure 1. Disclosed training hardware, by lab and chip platform



Data source: Hugging Face

At any point in time, the majority of signals are found for NVIDIA (1,581 out of 3,876), whereas at least 40% of top models do not disclose signs of the hardware used that we could find (Figure 2). For comparison, Epoch only reports training hardware for 38% of closed models in their notable models database, which means the share of closed-model developers directly reporting this will be even lower.

Figure 2. Which chips power the top AI models?

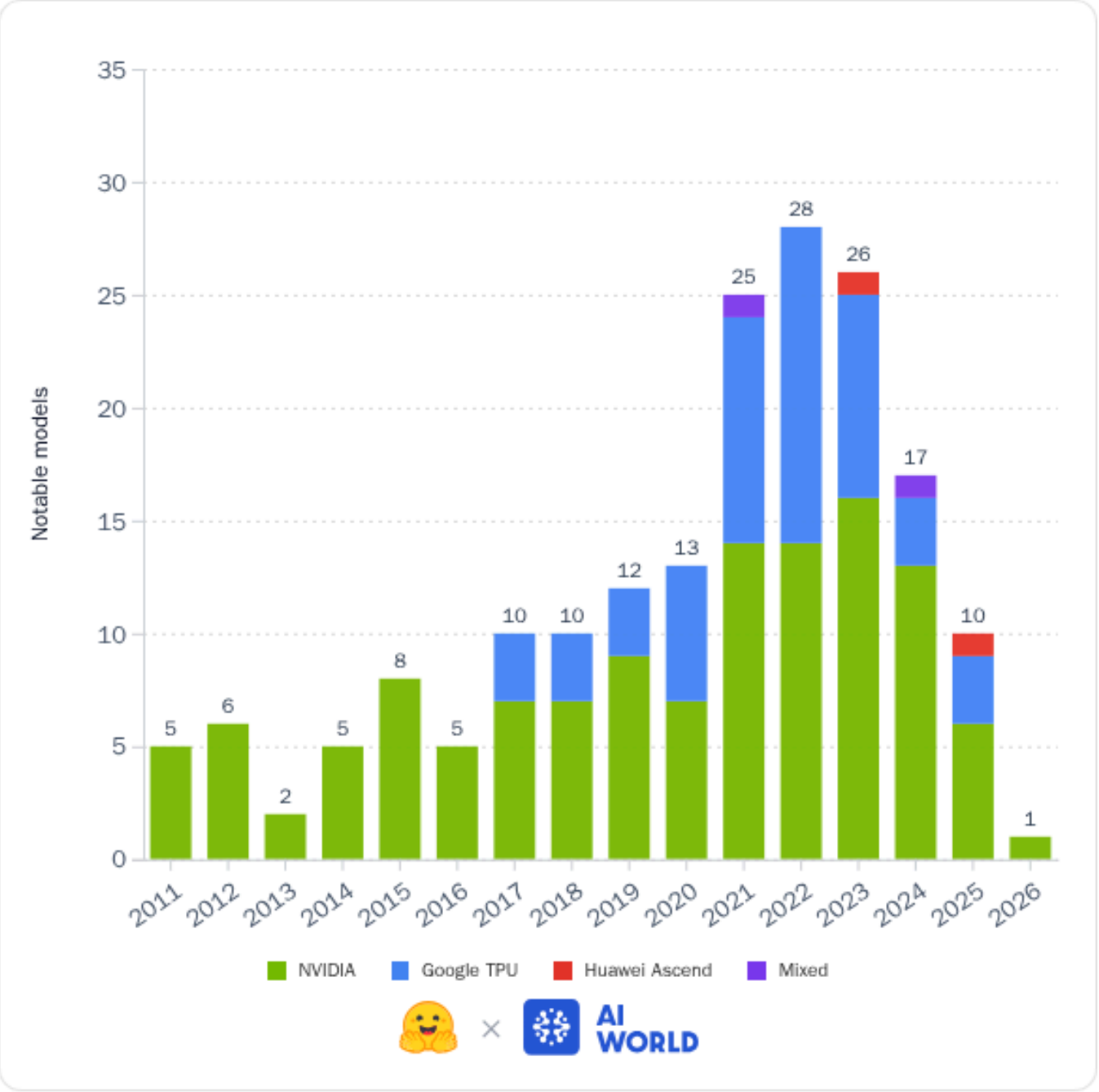


Data source: Hugging Face

Additional longitudinal data from Epoch's Notable Models suggests similar patterns for proprietary models (see Figure 3 below). Unsurprisingly, NVIDIA has held a central role as a hardware provider for some time. Google's TPU v2 chip, released in 2017, introduced Google into the provider market. Since 2023, Huawei's Ascend series also features, as do mixed

training approaches including the Amazon Trainium and NVIDIA hardware for training Amazon Nova Pro, and Huawei Ascend and NVIDIA for ERNIE 3.0 Titan.

Figure 3. Proprietary Notable AI Models by disclosed training-chip provider over time



Data source: Epoch AI

Data on which developers use which training hardware is relatively scarce. That said, based on publicly available information, the research team was able to assemble an initial map of the model training environment for open weight models. As displayed in Figure 4 below (bubbles are sized by the amount of links), most US-based firms leverage NVIDIA training chip platforms, with Google using their own in-house hardware together with Meta using a combination of Google and AMD chips. Anthropic and OpenAI have each been publicly associated with NVIDIA, TPUs, and Amazon Trainium across different model generations, but neither discloses which specific

hardware was used to train which model, nor the proportional split across providers. The list we can assemble from public sources is therefore a best effort floor estimate.

Figure 4. A curated map of AI labs and their training-chip platforms

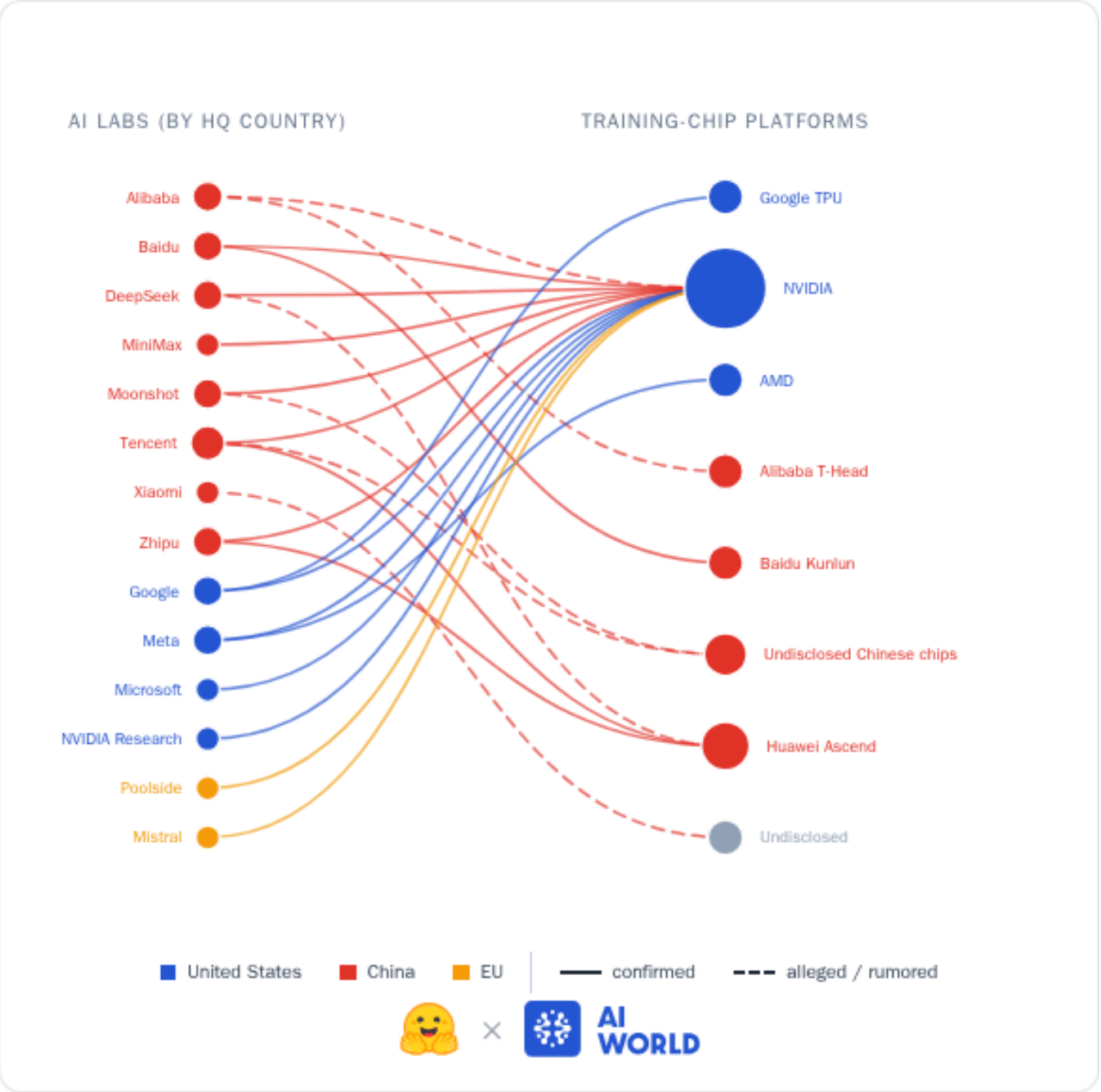


Figure 4 · Only lab–chip links with at least one citable public source about training hardware details are shown. This is therefore a best-effort floor estimate.

Data source: Various reports

For Chinese AI labs, however, our analysis indicates that their training hardware set-ups come from a more diverse set of both US and Chinese providers. Reports suggest that major model developers such as DeepSeek leverage both NVIDIA and Huawei's Ascend chips for model training. DeepSeek's attempt to train on Chinese chips is reflected in the fact that DeepSeek

V3 was trained exclusively on NVIDIA hardware while reports suggest the more recent V4 model was trained on a mix of NVIDIA and Huawei chips as well as adapted for strong performance on Huawei hardware. Additionally, per our analysis reports suggest that Alibaba also leverages NVIDIA chips as well as their in-house T-Head series. Accordingly, this information indicates that while Chinese firms are pushing towards using domestic training hardware, NVIDIA retains its central position in the model training ecosystem. Other western open developers with strong European ties, such as Mistral AI (through Coreweave) and Poolside, reportedly use Nvidia hardware as well. On the inference side specifically, DeepSeek's V4 release is reportedly served on Huawei Ascend 950 hardware, and its API pricing sits well below comparable US-served models. Whether the cost advantage comes from chip economics, vertical integration, or subsidized capacity is unclear from public information, but the signal that a frontier-tier open-weight model is being served at scale on domestic Chinese hardware is a meaningful shift from a year ago.

We colorized Epoch's graph on ML hardware based on the origin of the manufacturer, which suggests that improvement continues apace in both China and the United States. Top end performance is still achieved by US companies, indicated by blue dots in Figure 5, and is largely dominated by high-end NVIDIA chips. Moreover, despite Chinese efforts to indigenize hardware production, trend lines specific to the US (blue dash) and China (red dash) suggest that US chip performance increases continue at a faster rate. While hardware capabilities remain an important metric, as we highlight in the following section, performance differences in training and inference hardware demonstrates that the story is more multi-faceted than a focus on top-end capabilities alone would suggest.

Figure 5. Peak ML hardware performance over time, by chip country

Data source: Epoch AI

Inference Hardware Disclosure is as important as Training Hardware

Global model development is ongoing across diverse hardware environments. Yet, training and inference do not necessarily occur using the same hardware. As we discuss in this section, model performance can widely vary based on differences in training and inference hardware. Without improved transparency, cross-platform discrepancies will remain hard to diagnose,

undermining efforts to both evaluate model performance and assess global hardware dependencies driving political and policy calculations.

For instance, Epoch AI's benchmarking work found that GLM-4.6's accuracy on GPQA Diamond ranged from roughly 0% to 85% depending solely on the API provider. With the model itself held constant, that gap comes down to differences in each provider's inference setup — the underlying hardware included (see the appendix).

Much of this variation is rooted in the hardware itself. Floating-point addition is non-associative — $(A + (B + C)) \neq ((A + B) + C)$ — so moving a model across architectures (e.g., Huawei DaVinci/CANN to Nvidia CUDA) changes the order in which threads reduce operations and shifts the result LLM-42. Compounded across attention layers, these microscopic differences cause distinct behavioral divergence — enough that a model's outputs can be used to fingerprint the serving hardware. Other studies quantify the cost: up to 9% variation in accuracy and 9,000-token swings in response length tied in part to inference hardware study, and degraded performance when a model trained on one platform is served on another findings. The rest of the inference stack contributes too — even the backend alone can move results with hardware held constant — but, like hardware, it is seldom disclosed.

These findings have important implications. In terms of evaluating model performance, they suggest that benchmark metrics cannot be interpreted in isolation. A reported accuracy figure reflects not just model capability but also the specific hardware and inference configuration in which testing occurred. This complicates model evaluation and introduces noise when making capability assessments. Improving transparency in reporting training hardware can therefore help the technical community better understand sources of performance inefficiencies.

Evaluation capacity and transparency of hardware configuration is especially important for local inference: while token costs surge and GPUs are in high-demand, coders and heavy users especially are turning to local inference as well, owning a variety of hardware. Spinning up Llama.cpp to run models needs one line of code through the CLI, Claude code can be run with local models such as from Ollama, and even open source agent harnesses like Hermes from Nous Research are rapidly improving. In particular, running MLX for training and inference on Mac Apple Silicon chips has become an accessible and convenient way to run local AI with a Macbook. Performance inefficiencies in local inference may undermine these trends which enable users to deploy high quality, smaller, models for lower costs.

Additionally, without transparent insight into model hardware configurations, underlying global hardware dependencies in the AI ecosystem will remain difficult to assess in full. While our analysis above maps the training environment based on publically available data, many developers in both closed and open source environments do not disclose this information. For

the policy community, this makes it difficult to establish the baselines necessary to make informed, data driven, decisions.

Conclusions and Future Work

The findings laid out here suggest that: 1) there is an ongoing transparency deficit with respect to hardware disclosure, 2) that model training choices have diversified in the last year, and 3) that inferences choices have widened - particularly in the context of locally deployable models.

These results have both technical and political implications. In technical terms, an inability to reliably link training and inference environments means that evaluators cannot distinguish model performance issues from hardware-related inefficiencies. With respect to global political trends, a more transparent map of global hardware can help policymakers assess dependencies within the AI ecosystem and make decisions based on more reliable empirics rather than political narratives alone.

Finally, in the future we suggest further investigation into hardware dynamics with respect to inference. Specifically, the implications of increased 0-day support for Chinese suppliers, as well as the quickly expanding community of users leveraging local inference. Both factors remain understudied, yet warrant further attention.

Appendix

Methodology for tracking training hardware disclosure on the Hugging Face platform

The system tracks chip providers like Nvidia, AMD, Intel, Apple, Amazon and Google TPU. As well as Chinese providers like Huawei can become Baidu. Some known but untracked providers include Cerebras with their Cerebras-GPT family trained on their own chips, and Sambanova, which chips were used for training BLOOMChat.

The pipeline is a multi stage classifier ingesting metadata from three independent sources: the HF model card, the linked arXiv paper if it exists, and the GitHub repository associated with the model again if it exists. Then each is run through a dedicated classifier that extracts chip-related signals; strong evidence like using CUDA can short-circuit the pipeline and weaker

signals trigger a fallback call to an LLM which makes a committal judgment. The specific chip providers we track are in the tables below. The first graphic shows the abundance of NVIDIA compared to other chips providers, as well as the large share of models for which the hardware it was trained on was not disclosed or we couldn't capture.

Table A-1: Tracked (Western)

Provider	Key signals (strong)	Hardware SKU
nvidia	cuda, nvidia-smi, nvidia/cuda, nvidia-apex, EC2 p3/p4d/p5 families	A100, H100,
amd	rocm, hipify, rccl, amd-gpu, EC2 g4ad	MI250, MI30
intel	intel-extension-for-pytorch, gaudi, habana, openvino, neural-compressor	—
google_tpu	TPU, libtpu, cloud-tpu, TPUStrategy, jax.distributed, xm.xla_device	tpu-v2..v5
apple	MLX, coreml/coremltools, apple-silicon, anekit, qualified Metal/MPS	M1–M4 Pro/M
aws	trainium, inferentia, neuron-sdk, torch-neuronx, neuronx-cc	trn1/trn1n, in
qualcomm	QNN, SNPE, snapdragon, Qualcomm Hexagon, Hexagon DSP/NN/SDK	—

Table A-2: Tracked (Chinese)

Provider	Key signals (strong)
huawei_ascend	ascend, mindspore, cann, hccl, npu-smi, 昇腾, ASCEND_RT_VISIBLE_DEVICES, /de
cambricon	cambricon, cnml, cnnl, cndrv, bangpy, MLUDevice, torch-mlu
baidu_kunlun	kunlunxin, 昆仑(芯/核), xpurt, paddle-xpu, XPU_VISIBLE_DEVICES, baidu-xpu
moore_threads	MUSA, mthreads, moore-threads, torch-musa, musart, mccl, MUSA_VISIBLE_DEVIC
iluvatar	iluvatar, 天数智芯, ixrt, corex, ixsmi
hygon	hygon, 海光, hy-smi, hygon-dcu, hygon-dtk, DCU_VISIBLE_DEVICES
metax	metax, muxi, 沐曦, mxmaca, mx-smi, METAX_VISIBLE_DEVICES

To support reproducibility and scrutiny, both the code and the data behind this analysis are openly available. The full classification pipeline, including the source code and prompts used to extract hardware signals, lives in the project repository, and the captured dataset of model–chip associations underlying these figures is published as a Hugging Face dataset.

Author Contributions

CONCEPTUALIZATION	R. Praas, A. Ghosh
DATA CURATION	F. Ríos, R. Praas
INVESTIGATION	I. Reynolds, A. Ghosh, R. Praas
METHODOLOGY	F. Ríos, R. Praas, A. Ghosh
SOFTWARE	F. Ríos
VALIDATION	F. Ríos, I. Reynolds, R. Praas, A. Ghosh
FORMAL ANALYSIS	I. Reynolds, A. Ghosh, R. Praas
WRITING (ORIGINAL DRAFT)	F. Ríos, I. Reynolds, A. Ghosh, R. Praas
WRITING (REVIEW & EDITING)	F. Ríos, I. Reynolds, A. Ghosh, R. Praas, I. Solaiman
VISUALIZATION	F. Ríos
SUPERVISION	A. Ghosh, I. Solaiman

Citation

For attribution in academic contexts, please cite this work as

```
Francisco Ríos, Ian Reynolds, Robert Praas, Avijit Ghosh, Irene Solaiman (2026). "AI Hardware
Choices are Highly Variable and Sparsely Disclosed".
```

BibTeX citation

```
@misc{ríos2026_ai_hardware_choices_are_highly_variable_and_sparsely_disclosed,
  title={AI Hardware Choices are Highly Variable and Sparsely Disclosed},
  author={Francisco Ríos and Ian Reynolds and Robert Praas and Avijit Ghosh and Irene Solaiman},
  year={2026},

}
```

Reuse

References

1. Lawler, R. (2026). *Anthropic is in talks to use Microsoft’s AI chips too*. The Verge.

2. Emberson, L., & Sevilla, J. (2026). *Is a compute crunch coming?* Epoch AI.

3. Fedasiuk, R. (2026). *America’s Technology Siege Is Working as Intended*. Asia Society, Center for China Analysis.

4. Qu, T. (2026). *Alibaba Ramps Up AI Push With New Chip, Model Upgrades*. The Wall Street Journal.
5. Financial Times. (2026). *Huawei's AI chip sales surge as Nvidia stalls in China*. Financial Times.
6. Gerut, A. (2026). *Encrypted texts reveal how Nvidia chips and U.S. tech are being smuggled to China and Russia*. Fortune.
7. Google. (2026). *Our eighth generation TPUs: two chips for the agentic era*. Google.
8. Los Angeles Times. (2026). *Inside the AI compute crunch driving Google researchers to quit*. Los Angeles Times.
9. Shilov, A. (2025). *DeepSeek reportedly urged by Chinese authorities to train new model on Huawei hardware*. Tom's Hardware.
10. Reuters. (2026). *DeepSeek unveils new AI model tailored for Huawei chips as China pushes for tech autonomy*. Reuters.
11. Poolside. (n.d.). *Models*. Poolside.
12. CoreWeave. (2024). *Mistral AI and CoreWeave Demonstrate Partnership at NVIDIA GTC, Mistral AI Hackathon*. CoreWeave.
13. r/LocalLLaMA. (2026). *Buried lede: Deepseek v4 Flash is incredibly inexpensive from the official API for its weight category* [Online forum post]. Reddit.
14. Brand, F., & Denain, J.-S. (2025). *Why benchmarking is hard*. Epoch AI.
15. Yuan, J., et al. (2025). *Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference*. Hugging Face Papers: 2506.09501.
16. Qi, P., et al. (2025). *Defeating the Training-Inference Mismatch via FP16*. Hugging Face Papers: 2510.26788.
17. Gond, R., et al. (2026). *LLM-42: Enabling Determinism in LLM Inference with Verified Speculation*. Hugging Face Papers: 2601.17768.
18. Pape, D., Evertz, J., & Schönherr, L. (2026). *The Silent Hyperparameter: Quantifying the Impact of Inference Backends on LLM Reproducibility*. Hugging Face Papers: 2605.19537.
19. Angelo, J. (2026). *Microsoft reports are exposing AI's real cost problem: using the tech is more expensive than paying human employees*. Fortune.
20. Midha, A. [@anjneymidha]. (2026). *Post on X* [Post]. X.
21. Hugging Face. (n.d.). *Hardware*. Hugging Face.
22. SambaNova. (2023). *Introducing BLOOMChat-176B: the multilingual chat-based LLM*. SambaNova.