Models don’t have passports:

AI lineage crosses organizational and geographic borders

Amy Chang — Cisco Systems, Inc.
Manish Shah — VAIL, Inc.
Jonah Leshin — VAIL, Inc.
Ankit Garg — Cisco Systems, Inc.

AUGUST 2026

Table of contents

Executive summary

AI models are classified along multiple dimensions, including developer, model family, task and modality, access to weights, license, capability, and risk. In geopolitical and policy discussions, a model's publisher and the publisher's country of domicile are frequently used as shorthand for its origin. Those facts matter for accountability, legal jurisdiction, and operational risk, but they do not provide a complete account of how a model was produced. Modern models can inherit base weights, distillation teachers, synthetic training data, reward signals, tokenizers, and other components from models developed by different organizations and across different jurisdictions. We describe this condition as provenance entanglement: the technical lineage and dependencies of a model may cross organizational and national boundaries despite what is suggested by its release label.

This paper examines whether relationships of this kind remain detectable after post-training and release by another publisher. We study NVIDIA's Nemotron family and Alibaba's Qwen family because public documentation identifies Qwen base weights in the lineage of some Nemotron releases, while other Nemotron models use NVIDIA-developed or Meta Llama base weights. We apply two independently developed methods that observe different surfaces. Cisco's Model Provenance Kit (MPK) produces artifact-level signals from architecture metadata and weight-derived signals; this study uses its weight-derived identity score to construct model neighborhoods. VAIL compares behavioral fingerprints derived from model inference.

The results are directionally consistent across the two methods. In the 184-model MPK catalog, Qwen models account for 20.9% of the neighborhoods of 22 Nemotron references with documented Qwen base-weight ancestry, compared with a 12.0% catalog base rate (1.74× enrichment). Across all 37 Nemotron references, the observed enrichment is positive but less conclusive (1.27×; 95% CI: 0.99×–1.56×).

VAIL's 1,159-model catalog produces the same directional cross-family pattern. Qwen models account for 21.7% of the neighborhoods of all 46 Nemotron models, compared with a 14.9% catalog base rate (1.46× enrichment). For the 27 Nemotron models with documented Qwen base-weight ancestry, the share rises to 28.1% (1.89×). Within Nemotron, mean Qwen-neighbor shares follow the documented ordering: Qwen-base-weight models at 28.1%, NVIDIA-native models at 20.4%, and Llama-base-weight models at 3.9%. The corresponding MPK means are 20.9%, 11.9%, and 0.0%. One-sided permutation tests support the ordered association under both methods (MPK p=0.00001; VAIL p=0.00002).

These findings are evidence of concordance with documented lineage, not an independent reconstruction of causal training history. Similarity can also arise from shared architectures, tokenizers, data, teachers, or training objectives. The analysis does not establish the geographic origin of training data, continuing access by an upstream provider, or whether a deployment is secure.

Country-of-origin classifications often serve as proxies for a range of risks that organizations seek to control, but those risks implicate technical and operational factors that may not be captured by one geographic label alone. Cisco and VAIL research does not determine which restrictions governments or organizations should adopt; rather, it demonstrates that the evidence required depends on the underlying policy or regulatory objective. Rules addressing inherited vulnerabilities or behavioral traits could focus on weight lineage, training dependencies, artifact inspection, and behavioral evaluation. Procurement and assurance processes could require a model bill of materials and combine disclosure with proportionate technical verification. Publisher domicile remains relevant to jurisdiction and accountability, but it does not by itself establish technical independence.

01 Background and motivation

AI models are described and classified along several overlapping dimensions. Common dimensions include the developer or publisher; model family and base model; task and input–output modality; access to weights and licensing terms; scale and capabilities; intended use; and regulatory or operational risk. Model repositories encode many of these dimensions through model-card metadata, while governance frameworks distinguish categories such as general-purpose models and general-purpose models with systemic risk (Mitchell et al. 2019; OECD.AI 2026; European Parliament and Council of the European Union 2024).

Publisher identity and domicile are nevertheless prominent in geopolitical, industrial-policy, and national-security discussions, where models are often grouped as, for example, U.S.- or Chinese-developed and then further described as open- or closed-weight (Rahman 2024; Meinhardt et al. 2025). These labels are useful for attribution and accountability, but they are incomplete proxies for a model's technical origin, dependencies, and behavior. A publisher identifies who released an artifact; it does not necessarily identify everything from which that artifact was derived.

A model's supply chain is better represented by a graph than by a single country-of-origin label. That graph may include pretraining and post-training datasets, base-model weights, models used to generate synthetic data, formal distillation teachers, reward models, merged checkpoints, training code and infrastructure, and the organizations that developed, hosted, or modified these components. Prior work similarly characterizes foundation models as parts of a larger ecosystem of interdependent datasets, models, organizations, and applications (Bommasani, Soylu, et al. 2023). Documentation mechanisms such as model cards and datasheets were developed to make portions of these relationships visible (Mitchell et al. 2019; Gebru et al. 2021). Nevertheless, disclosures about upstream resources and model-development practices remain incomplete and uneven (Bommasani, Klyman, et al. 2023).

We use *model provenance* in the narrower sense specified by the Model Provenance Constitution: the verifiable derivation history of a model's trained weights (Cisco AI Defense 2026a; Aghaei et al. 2026). Under this definition, two models are provenance-linked only when a causal weight-derivation chain connects them through direct checkpoint descent, formal knowledge distillation, mechanical transformation, identity, or a transitive composition of those relationships. Shared architecture, tokenizer, training data, family name, organization, benchmark performance, or behavioral similarity do not establish model provenance when taken alone.

Terminology is not uniform across the broader literature: governance frameworks also discuss data provenance, organizational accountability, and third-party value-chain dependencies (Bommasani, Soylu, et al. 2023; Autio et al. 2024). To avoid conflating these uses, we reserve the unqualified term model provenance for causal trained-weight derivation and use qualified terms for other supply-chain facts.

01 /
Background and motivation

Narrowly defined model provenance sits within a broader model supply chain.
For governance purposes, at least six categories of supply-chain information could be distinguished:

1. **Organizational attribution**
   the entities that develop, release, license, or distribute a model;

2. **Model-weight provenance**
   the causal chain of base checkpoints, fine-tunes, formal distillation, merges, pruning operations, and other transformations of trained parameters;

3. **Data provenance**
   the sources and transformations of pretraining, post-training, preference, and evaluation data;

4. **Model-mediated training influence**
   models used to generate synthetic data, labels, critiques, reasoning traces, or reward signals, noting that formal distillation is provenance-linked under the Constitution while synthetic-data training alone currently is not;

5. **Operational control**
   the entities that host, update, or retain technical authority over a deployed system; and

6. **Behavioral characteristics**
   empirically observable similarities that may supply evidence about, but do not themselves define, a provenance relationship.

This supply-chain taxonomy defines the broader governance problem; the present study does not attempt to measure all six categories. We do directly examine two observable surfaces: artifact-level similarity related principally to model-weight provenance, and inference-level behavioral similarity. We use public release documentation to assign publisher and disclosed base-weight lineage. We do not independently measure the geographic origin of training data, attribute behavioral similarity to particular teacher models, or assess hosting, telemetry, update authority, and other forms of continuing operational control.

The distinction matters because downstream modification cannot be assumed to erase every upstream property. Prior work shows that pretrained weights can be intentionally modified so that backdoor behaviors survive downstream fine-tuning (Kurita et al. 2020). Deliberately constructed deceptive or trigger-dependent behaviors can persist through supervised fine-tuning, reinforcement learning, and adversarial training (Hubinger et al. 2024), and specially constructed backdoors can transfer from teacher to student models through knowledge distillation (Cheng et al. 2024). These studies do not establish that all upstream traits survive all transformations. They establish the narrower security point that a new publisher label or a new round of post-training does not, by itself, demonstrate independence from upstream artifacts.

This motivates ex-post technical analysis. Self-reported documentation remains essential, but artifact and behavioral fingerprints may provide additional detection evidence when documentation is incomplete or when an organization needs to test whether an observed model is consistent with a claimed lineage. The appropriate claim is evidentiary rather than deterministic. Provenance is a factual property of the causal derivation chain; a fingerprint is a detection signal. Similarity may motivate or corroborate an inference of lineage, but it does not by itself satisfy the Constitution's evidence standard or prove a unique causal training history.

02 Contribution and scope

This study makes three contributions:

First

We situate narrowly defined model-weight provenance within the broader, multidimensional model supply chain and distinguish causal weight lineage from publisher identity, data provenance, model-mediated training influence, behavioral resemblance, and operational control. This distinction clarifies why a model released by a provider in one jurisdiction may retain material dependencies on models developed elsewhere, while a foreign-published open-weight artifact may be deployed without continuing access or control by its original publisher.

Second

We test whether two independently developed fingerprinting methods provide convergent ex-post evidence consistent with documented technical lineage. Cisco's Model Provenance Kit compares model artifacts using architecture metadata, tokenizer characteristics, and weight-derived signals (Cisco AI Defense 2026b). VAIL compares models using behavioral fingerprints derived from inference (VAIL Inc. 2026). The methods observe different surfaces of a model and therefore supply complementary, not interchangeable, evidence.

Third

We present a case study of NVIDIA Nemotron and Alibaba Qwen. Nemotron is particularly informative because its releases include multiple documented base-weight relationships. Some Nemotron models are post-trained from Qwen checkpoints, some use NVIDIA-developed base weights, and others descend from Meta Llama checkpoints. NVIDIA also documents uses of Qwen models in synthetic data pipelines (Nguyen et al. 2025; NVIDIA 2025). This case therefore illustrates why base-weight ancestry, model-mediated training influence, and publisher identity must be kept conceptually separate.

Our empirical claims are deliberately limited to whether artifact and behavioral similarity align with documented base-weight groupings. We do not reconstruct complete training pipelines, verify all disclosures, infer the nationality of training data, determine whether a foreign entity retains access to a deployment, or treat similarity as proof of direct derivation.

03 Methods

EVIDENTIARY DESIGN

The study combines three evidentiary layers. *Documented derivation* comes from model cards and release documentation that explicitly identify a parent checkpoint or derivation mechanism. *Artifact-level detection evidence* comes from Cisco's Model Provenance Kit. *Behavioral evidence* comes from VAIL's behavioral fingerprints. The analysis asks whether the two measured layers are consistent with the relationships reported in the first. It does not treat model names as provenance evidence, documentation as a complete account of every upstream dependency, or fingerprints as a complete provenance graph.

This design follows the Constitution's distinction between a provenance relationship and a provenance-detection signal. Under its evidence standard, a pairwise provenance label requires official documentation naming the parent and mechanism, checkpoint verification, or authoritative third-party analysis (Cisco AI Defense 2026a). Statistical similarity, architecture, tokenizer overlap, and naming are insufficient on their own. Accordingly, documented base-weight ancestry determines the Nemotron tiers in this study; the fingerprints are evaluated for concordance with those labels. Aggregate neighborhood enrichment is not used to declare every Nemotron–Qwen pair provenance-linked.

FINGERPRINT CATALOGS

Cisco artifact fingerprints are generated with Cisco's Model Provenance Kit (MPK) version 1.1.0 and a fixed 184-model reference-catalog snapshot dated August 5, 2026. The toolkit compares architecture metadata, tokenizer structure and vocabulary overlap, and five weight-derived signals. The present analysis uses MPK's *identity score*, a weighted average of the five weight-derived signals, to rank all models in the catalog. This continuous, weight-only measure supports a consistent nearest-neighbor ordering; MPK's separately reported tokenizer score and metadata-gated operational scan score are not used in the neighborhood analyses (Cisco AI Defense 2026b). We treat the resulting similarities as *artifact-level detection evidence*, not as calibrated probabilities or proof of derivation.

VAIL fingerprints are generated with VAIL's technology (internal version 1.1; catalog snapshot dated July 27, 2026), described at Project VAIL. VAIL constructs fingerprints from model behavior during inference: the method runs a forward pass on fixed input tokens, extracts a linear approximation of the model's mapping from the resulting logits, and forms a vector from the magnitudes of the corresponding eigenvalues. Fingerprint vectors from two models are compared to produce a similarity score between 0 and 1. These are referred to here as *behavioral fingerprints.* (VAIL Inc. 2026)

The Cisco cross-family analysis uses an artifact-fingerprint catalog containing 184 reference models; the VAIL analysis uses fingerprints for 1,159 models. Both datasets cover publicly available, predominantly open-weight models. Because catalog membership and fingerprint construction differ, we compare directional cross-family patterns rather than absolute similarity values. The Cisco cross-family analysis uses the nearest-neighbor ordering computed across the full 184-model catalog using the identity score described above.

03 / Methods
Families and Nemotron base-weight tiers

FAMILIES AND NEMOTRON BASE-WEIGHT TIERS

"NVIDIA" denotes every NVIDIA-published model in the relevant comparison pool. "Nemotron" denotes the NVIDIA-published, Nemotron-branded language-model set used as the reference set. The roster includes one vision-language release, nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1; the present analysis does not separately attribute fingerprints to its language and vision components. Qwen, Google, and Meta families are assigned from the corresponding catalog metadata and model documentation.

For the within-Nemotron analysis, the 46 VAIL Nemotron reference models are partitioned by documented *base-weight ancestry:* 27 Qwen-base-weight models, 10 NVIDIA-native models, and nine Llama-base-weight models. These labels are mutually exclusive for the purposes of the test. They do not assert that a model has no other cross-family dependencies. In particular, an NVIDIA-native or Llama-base-weight model may have been trained on Qwen-generated synthetic data or may have used other models as teachers, judges, or reward models. The tier roster is reported in the appendix.

The MPK catalog's 37 Nemotron reference models are partitioned by the same documented base-weight ancestry: 22 Qwen-base-weight, nine NVIDIA-native, and six Llama-base-weight models. This roster is reported in the appendix.

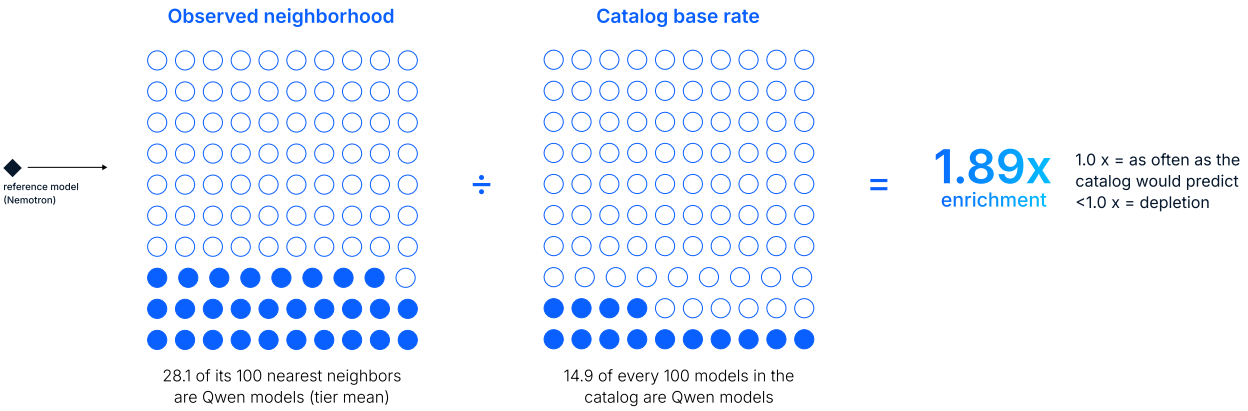


Figure 1. Construction of the neighborhood enrichment statistic. *The diamond marks a reference model; the grid is its 100 nearest neighbors, shaded where a neighbor belongs to the target family. Enrichment divides the mean neighbor share by that family's share of the comparison pool, so 1.0× is representation at the catalog base rate. Illustrated with the Qwen-base-weight Nemotron tier under VAIL fingerprints (n=27): 28.1% is the tier mean, and position within the grid carries no meaning.*

Consistent with the Constitution's evidence standard, a family name is used only to locate relevant documentation, not to establish a provenance label. Final tier assignment requires release documentation that explicitly identifies the parent checkpoint or derivation mechanism. A model supported only by naming convention would remain unclassified until checkpoint evidence or qualifying third-party analysis became available.

Google and Meta families are used as comparison families because the included models have no documented direct Qwen-weight ancestry. They are not treated as verified-independent controls: absence of a disclosed relationship does not establish absence of shared data, teachers, architectures, or training practices.

03 / Methods
Neighborhood enrichment / Hypotheses and uncertainty

NEIGHBORHOOD ENRICHMENT

For VAIL, a model's neighborhood consists of its closest 100 neighbors. For Cisco, it consists of its closest 15 neighbors. Excluding the reference model, these correspond to 8.6% of the VAIL comparison pool and 8.2% of the Cisco comparison pool.

Tables 1a and 1b use the notation *reference family→target family*. For example, Nemotron→Qwen calculates, for each Nemotron reference model, the fraction of its neighborhood occupied by Qwen models. The *observed share* is the average of those model-level fractions. The *base rate* is the target family's share of the full comparison pool, excluding the reference model itself. *Enrichment* is observed share divided by base rate: 1.0× indicates representation at the corpus base rate, values above 1.0× indicate enrichment, and values below 1.0× indicate depletion.

Neighborhood enrichment is directional as a statistic. Nemotron→Qwen asks a different question from Qwen→Nemotron because the reference sets, target base rates, and local neighborhoods differ. This statistical direction does not establish the direction of a provenance relationship. Where parent–child direction is reported, it comes from release documentation rather than the neighborhood calculation.

HYPOTHESES AND UNCERTAINTY

We evaluate two hypotheses. First, the *cross-family hypothesis* predicts that Qwen models will be overrepresented in Nemotron neighborhoods, particularly for the Qwen-base-weight tier, and that this pattern will exceed the comparison-family rows. Second, the *ordered-lineage hypothesis* predicts that model-level Qwen-neighbor shares will follow the ordering Qwen-base-weight > NVIDIA-native > Llama-base-weight.

Cross-family confidence intervals quantify uncertainty by resampling reference models with 10,000 bootstrap samples. These intervals describe variation across the reference models represented in each catalog. They do not account for uncertainty caused by catalog construction, family-label error, or the choice of neighborhood size.

The ordered-lineage hypothesis is evaluated separately in each catalog with a one-sided Jonckheere–Terpstra permutation test. For VAIL, we hold the 46 model-level Qwen-neighbor shares fixed and randomly reassign them among tiers while preserving tier sizes of 27, 10, and nine. For MPK, we apply the same procedure to the 37 model-level shares while preserving tier sizes of 22, nine, and six. Each test uses 100,000 permutations. The permutation p-value is calculated as $(r+1)/100{,}001$, where r is the number of permuted statistics at least as large as the observed statistic.

04 Results

CROSS-FAMILY ENRICHMENT

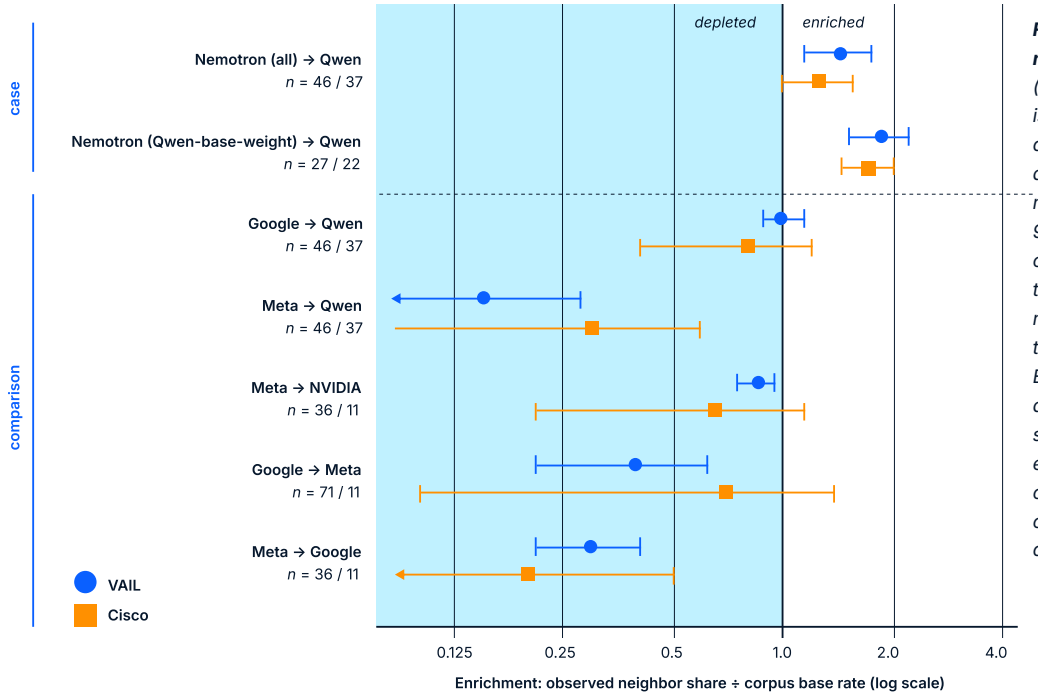


Figure 2. Cross-family neighborhood enrichment *(Tables 1a and 1b). Enrichment is the observed neighbor share divided by the corpus base rate, on a logarithmic scale; the rule at 1.0× marks the base rate. Error bars are 95% bootstrap confidence intervals over reference models, and n gives the VAIL and Cisco reference-model counts. Intervals reaching the left boundary are truncated. Enrichment places the results on a common base-rate-normalized scale, but the values are not directly equivalent effect sizes because catalog membership, fingerprint construction, and neighborhood definitions differ.*

Figure 2 and Tables 1a and 1b report the cross-family analysis using VAIL behavioral fingerprints and Cisco artifact fingerprints, respectively.

Table 1a. Cross-family enrichment using VAIL behavioral fingerprints

Reference → target	*n* (ref)	Observed	Base	Enrichment	95% CI
Nemotron (all) → Qwen	46	21.7%	14.9%	1.46×	[1.15×, 1.76×]
Nemotron (Qwen-base-weight) → Qwen	27	28.1%	14.9%	1.89×	[1.53×, 2.24×]
Google → Meta	71	1.2%	3.1%	0.40×	[0.21×, 0.62×]
Meta → Google	36	1.8%	6.1%	0.30×	[0.21×, 0.40×]
Google → Qwen	71	14.9%	14.9%	1.00×	[0.88×, 1.14×]
Meta → Qwen	36	2.2%	14.9%	0.15×	[0.06×, 0.28×]
Meta → NVIDIA	36	5.1%	6.0%	0.86×	[0.76×, 0.96×]

In the VAIL catalog, Qwen models account for 21.7% of the neighborhoods of all 46 Nemotron references, compared with a 14.9% corpus base rate. This corresponds to 1.46× enrichment (95% CI: 1.15×–1.76×). Restricting the reference set to the 27 Qwen-base-weight Nemotron models increases the observed share to 28.1% and the enrichment to 1.89× (95% CI: 1.53×–2.24×).

04 / Results
Cross-family enrichment

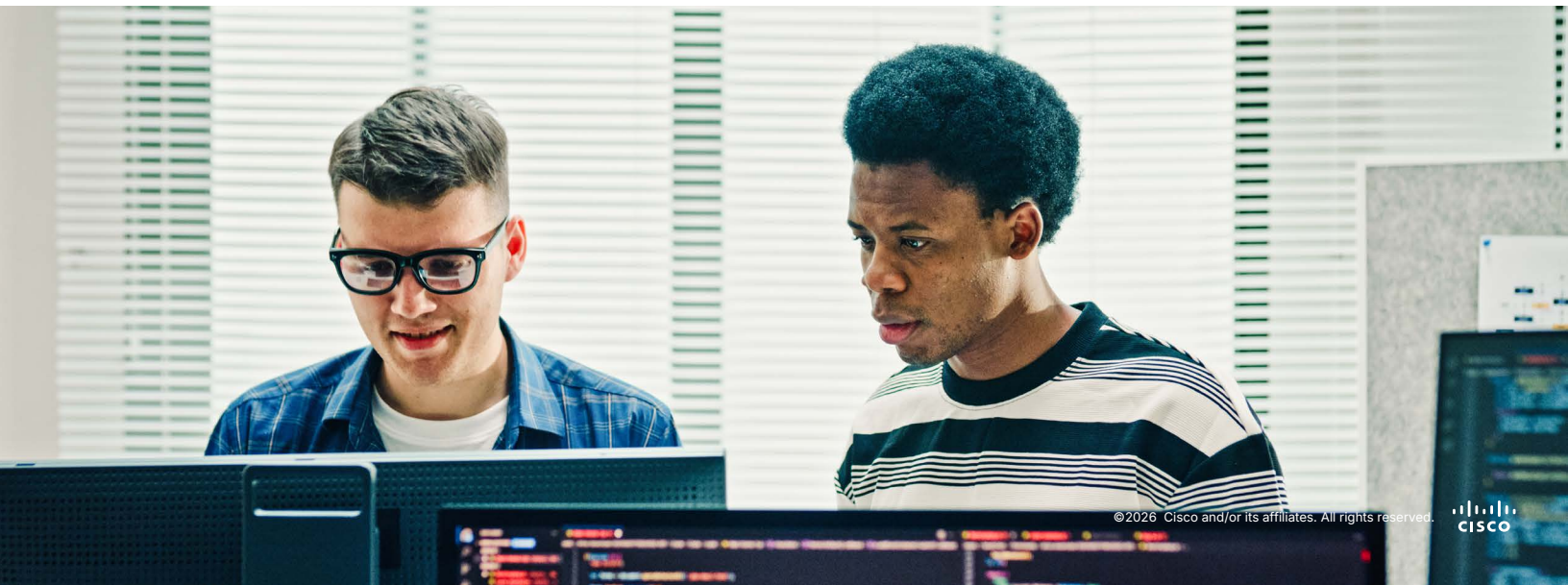
The comparison-family rows do not show the same pattern. Google → Qwen occurs at the Qwen base rate, while Meta → Qwen and both Google–Meta directions are depleted. Thus, the VAIL result is not explained solely by Qwen's prevalence in the catalog.

Table 1b. Cross-family enrichment using Cisco artifact fingerprints

Reference → target	*n* (ref)	Observed	Base	Enrichment	95% CI
Nemotron (all) → Qwen	37	15.3%	12.0%	1.27×	[0.99×, 1.56×]
Nemotron (Qwen-base-weight) → Qwen	22	20.9%	12.0%	1.74×	[1.46×, 2.02×]
Google → Meta	11	4.2%	6%	0.71×	[0.10×, 1.41×]
Meta → Google	11	1.2%	6%	0.20×	[0.00×, 0.50×]
Google → Qwen	11	9.7%	12%	0.81×	[0.40×, 1.21×]
Meta → Qwen	11	3.6%	12%	0.30×	[0.00×, 0.60×]
Meta → NVIDIA	11	13.3%	20.2%	0.66×	[0.21×, 1.17×]

The Cisco catalog yields the same directional result. Across 37 Nemotron reference models, the Qwen-neighbor share is 15.3%, corresponding to 1.27× enrichment (95% CI: 0.99×–1.56×). For the 22 Qwen-base-weight Nemotron models, the observed share is 20.9%, or 1.74× enrichment (95% CI: 1.46×–2.02×). The Google and Meta comparison rows range from depleted to statistically compatible with the base rate.

The Cisco and VAIL analyses use similarly sized Nemotron reference sets: 37 models under MPK and 46 under VAIL. Both show Qwen-neighbor enrichment within the documented Qwen-base-weight tier (1.74× for MPK and 1.89× for VAIL), with confidence intervals above 1.0×. Evidence for the pooled Nemotron set is weaker under MPK: the observed enrichment is 1.27×, but its confidence interval includes the base-rate value (95% CI: 0.99×–1.56×). The tier-specific result is therefore the strongest point of agreement between the methods.



04 / Results
Nemotron base-weight tiers

NEMOTRON BASE-WEIGHT TIERS

The second analysis asks whether Qwen-neighbor shares follow the documented base-weight ordering within the Nemotron reference models, under both VAIL and Cisco artifact fingerprints. Table 2a reports the model-level Qwen-neighbor shares for the 46 VAIL Nemotron reference models; Table 2b reports the same statistic for the 37 Cisco Nemotron reference models.

Table 2a. VAIL Nemotron base-weight tiers: Qwen share of top-100

Tier	*n*	Mean Qwen %	Median	Range	Enrichment vs. base
Qwen-base-weight	27	28.1%	25%	0–54%	1.89x
NVIDIA → native	10	20.4%	21%	1–38%	1.37x
Llama-base-weight	9	3.9%	1%	0–18%	0.26×

Table 2b. MPK Nemotron base-weight tiers: Qwen share of top-15

Tier	*n*	Mean Qwen %	Median	Range	Enrichment vs. base
Qwen-base-weight	22	20.9%	20%	7–33%	1.74x
NVIDIA → native	9	11.9%	7%	7–33%	0.99x
Llama-base-weight	6	0.0%	0%	0–0%	0.00×

Both catalogs show the observed means following the hypothesized ordering. Under VAIL, Qwen-base-weight models have a mean Qwen-neighbor share of 28.1%, followed by NVIDIA-native models at 20.4% and Llama-base-weight models at 3.9% (Jonckheere–Terpstra J=480, one-sided permutation p=0.00002). Under Cisco artifact fingerprints, Qwen-base-weight models have a mean Qwen-neighbor share of 20.9%, followed by NVIDIA-native models at 11.9% and Llama-base-weight models at 0.0% (J=347, one-sided permutation p=0.00001). Both tests evaluate the ordering of model-level neighborhood shares; neither tests whether each tier is enriched relative to the catalog base rate.

The methods differ most clearly for the NVIDIA-native tier. Its mean Qwen-neighbor share is approximately at the catalog base rate under MPK (11.9% observed against a 12.0% base rate) but above it under VAIL (20.4% against 14.9%). One possible interpretation is that behavioral similarity reflects training-data or teacher influence not accompanied by Qwen weight ancestry, while the artifact method is more sensitive to weight-level relationships. The present analysis cannot distinguish that explanation from differences in catalog composition, fingerprint construction, or sampling variation. The result supports an ordered association with documented tier labels, not the conclusion that Qwen base weights caused every observed behavioral resemblance.

05 Discussion

WHAT THE RESULTS ESTABLISH

Across both methods, Nemotron models with documented Qwen base-weight ancestry have Qwen-neighbor shares above the corresponding catalog base rates. The pooled Nemotron association is also positive in both catalogs, although the MPK confidence interval narrowly includes the base-rate value. Both methods recover the documented ordering of base-weight tiers. This convergence is notable because Cisco and VAIL do not measure the same object: Cisco analyzes properties of model artifacts, whereas VAIL analyzes behavior during inference.

These results provide ex-post detection evidence *consistent with* documented technical lineage. They show that publisher identity does not erase detectable relationships to upstream model families and that such relationships can remain visible after post-training and re-release under another publisher's model family. The documented parent checkpoints and mechanisms establish the tier labels; concordant fingerprints show that signals associated with those relationships remain detectable. The aggregate enrichment analysis does not itself create new pairwise provenance labels.

WHAT THE RESULTS DO NOT ESTABLISH

The study does not prove a unique derivation path. A neighborhood fingerprint is a similarity measure, not a causal reconstruction of training. Behavioral similarity may arise through inherited weights, synthetic data, distillation, teacher or reward models, shared tokenizers and architectures, common data, similar post-training objectives, or convergence induced by benchmarks. Artifact similarity narrows some of these possibilities but, under the Constitution, statistical weight similarity is still insufficient on its own to establish a causal provenance link. It also does not establish the geographic origin of training data or continuing control by an upstream provider.

Under the Constitution, a qualifying causal derivation either occurred or did not, whereas detectable similarity is continuous and may weaken after modification or across multiple hops. The fingerprints studied here estimate association; they do not redefine that provenance boundary.

The study also does not classify models as nationally secure or insecure. It does not examine whether Alibaba, NVIDIA, or another entity hosts a particular deployment; receives prompts or outputs; can update or disable the model; or is subject to legal process affecting the deployment. Those are questions of organizational attribution and operational control requiring different evidence.

05 / Discussion
Governance and policy implications

GOVERNANCE AND POLICY IMPLICATIONS

Policies that assign models a single country of origin based on publisher domicile collapse distinct risks into one label. Policymakers often default to country-of-origin rules as a way of containing or managing risk by controlling technical ancestry, inherited characteristics, or operational access. We find that publisher domicile alone may be an incomplete proxy for these endpoints. It can also be overinclusive when a locally hosted open-weight model is associated with a foreign publisher that has no continuing access to the deployment, no update channel, and no ability to receive user data.

This does not imply that publisher domicile is irrelevant or that a particular restriction is warranted. While domicile may itself be a direct criterion for sanctions, procurement preferences, jurisdiction, or industrial policy, the narrower technical point is that domicile and provenance describe different properties, and evidence of one does not establish the other. Lineage evidence can identify a pathway through which an upstream characteristic might persist, but it does not establish that a particular vulnerability, bias, or behavior was inherited or remains material in the downstream model. That question requires separate security and behavioral evaluation.

The findings therefore help distinguish the evidence relevant to different policy questions:

1. Policies addressing **remote access, data exfiltration, service interruption, or legal compulsion** turn on hosting, telemetry, update authority, ownership, and operational control.
2. Policies addressing **inherited vulnerabilities or behavior** implicate model-weight provenance, data provenance, model-mediated training influence, artifact inspection, and behavioral evaluation.
3. Assessments of **supply-chain dependency** concern critical upstream models, datasets, infrastructure, and licenses, together with substitution and continuity risks.
4. Policies addressing **sanctions, procurement preferences, or industrial strategy** may use organizational or jurisdictional criteria directly; doing so is analytically distinct from finding technical independence or dependence.

Supply-chain documentation provides one mechanism for keeping these categories separate. A model bill of materials, for example, can record base checkpoints, derivation mechanisms, merges, major datasets, synthetic-data generators, teacher and reward models, training services, licenses, and entities with post-deployment access or update authority. Whether and when such documentation is required is a policy choice. The U.S. National Institute of Standards and Technology (NIST) similarly treats models, datasets, and other third-party components as part of the generative-AI value chain (Autio et al. 2024), while the European Union AI Act imposes documentation duties on general-purpose model providers for the benefit of downstream integrators (European Parliament and Council of the European Union 2024).

Artifact and behavioral fingerprints can complement self-reported records by identifying inconsistencies, prioritizing models for deeper review, and providing detection evidence when full records are unavailable. Consistent with the Constitution's conservative evidence standard, similarity could trigger further inquiry, but does not automatically establish prohibited ancestry or determine the appropriate control. If a governance framework uses such evidence, possible responses may vary with the policy objective, deployment context, and strength of the evidence. This study does not evaluate the necessity or proportionality of any particular controls.

The relevant facts may also change over a model's lifecycle. Fine-tuning, merging, and quantization create additional provenance links, while retrieval augmentation, system prompts, hosting changes, and remote updates may alter behavior or operational dependencies without changing weight provenance. A publisher label is static; the relevant supply chain is not.

06 Limitations

First

Both analyses are conditional on their catalogs. Base rates, neighbor shares, and enrichment values are all defined relative to catalog composition, and both catalogs are curated selections rather than draws from a well-defined population of models. The agreement between two independently constructed catalogs suggests the directional pattern is not an artifact of either selection, but the specific magnitudes should be read as properties of the catalogs studied rather than precise estimates for open-weight models in general.

Second

Several references are variants from the same model-development series and may not represent fully independent development events. The model-level uncertainty estimates may therefore understate dependence among closely related releases.

Third

Family and base-weight tiers rely on public documentation. Documentation may be incomplete, ambiguous, or incorrect, and naming alone is not sufficient under the Constitution's evidence standard. The comparison families should be understood as having no *documented* direct Qwen-weight ancestry, not as proven-independent models.

07 Conclusion

Model provenance cannot be reduced to the name or domicile of the organization that releases a model. In the Nemotron–Qwen case, both weight-derived artifact similarity and behavioral similarity show cross-family associations consistent with documented technical lineage and recover the documented ordering of base-weight groups. Together, these findings demonstrate the value of combining disclosure with ex-post technical evidence.

They also demonstrate the limits of any single evidentiary signal. Model-weight provenance, data provenance, model-mediated training influence, publisher identity, operational control, and observed behavior answer different questions. The governance relevance of any one of these facts depends on the policy objective and should not be inferred from another dimension by proxy. Models do not have passports; they have supply chains.

Nemotron base-weight tier roster

Table 3 lists the 46 VAIL Nemotron models used in Table 2a base-weight tier analysis; Table 4 lists the 37 Cisco Nemotron models used in Table 2b base-weight tier analysis. The categories are mutually exclusive within each roster. One entry in Table 3 is a vision-language release; its tier refers to the documented language-model base-weight ancestry. The model sets for the cross-family analyses are not listed here: they use the full VAIL (1,159-model) and Cisco (184-model) catalogs, whose membership is defined by the catalogs themselves rather than a study-specific selection.

Table 3. Nemotron reference models by base-weight tier

Tier	Model
Qwen-base-weight	nvidia/AceMath-RL-Nemotron-7B
Qwen-base-weight	nvidia/AceReason-Nemotron-1.1-7B
Qwen-base-weight	nvidia/AceReason-Nemotron-14B
Qwen-base-weight	nvidia/Nemotron-Cascade-14B-Thinking
Qwen-base-weight	nvidia/Nemotron-Cascade-8B
Qwen-base-weight	nvidia/Nemotron-Cascade-8B-Thinking
Qwen-base-weight	nvidia/Nemotron-Orchestrator-8B
Qwen-base-weight	nvidia/Nemotron-Research-Reasoning-Qwen-1.5B
Qwen-base-weight	nvidia/Nemotron-Terminal-14B
Qwen-base-weight	nvidia/Nemotron-Terminal-32B
Qwen-base-weight	nvidia/Nemotron-Terminal-8B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-1.1-14B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-1.1-32B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-1.1-7B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-14B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-32B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-32B-IOI
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-7B
Qwen-base-weight	nvidia/OpenMath-Nemotron-1.5B
Qwen-base-weight	nvidia/OpenMath-Nemotron-14B
Qwen-base-weight	nvidia/OpenMath-Nemotron-14B-Kaggle
Qwen-base-weight	nvidia/OpenMath-Nemotron-32B
Qwen-base-weight	nvidia/OpenMath-Nemotron-7B

Nemotron base-weight tier roster

Tier	Model
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-1.5B
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-14B
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-32B
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-7B
NVIDIA-native	nvidia/Nemotron-4-Mini-Hindi-4B-Base
NVIDIA-native	nvidia/Nemotron-4-Mini-Hindi-4B-Instruct
NVIDIA-native	nvidia/Nemotron-Flash-1B
NVIDIA-native	nvidia/Nemotron-Flash-3B-Instruct
NVIDIA-native	nvidia/Nemotron-Mini-4B-Instruct
NVIDIA-native	nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-12B-v2
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-12B-v2-Base
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-9B-v2
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-9B-v2-Base
Llama-base-weight	nvidia/Llama-3.1-Nemotron-70B-Instruct-HF
Llama-base-weight	nvidia/Llama-3.1-Nemotron-8B-UltraLong-1M-Instruct
Llama-base-weight	nvidia/Llama-3.1-Nemotron-8B-UltraLong-2M-Instruct
Llama-base-weight	nvidia/Llama-3.1-Nemotron-8B-UltraLong-4M-Instruct
Llama-base-weight	nvidia/Llama-3.1-Nemotron-Nano-4B-v1.1
Llama-base-weight	nvidia/Llama-3.1-Nemotron-Nano-8B-v1
Llama-base-weight	nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1
Llama-base-weight	nvidia/Llama-3_3-Nemotron-Super-49B-v1
Llama-base-weight	nvidia/Llama-3_3-Nemotron-Super-49B-v1_5

MPK base-weight tier roster

Table 4 lists the 37 Nemotron models used in the Cisco MPK base-weight tier analysis (Table 2b), partitioned by documented base-weight ancestry. nvidia/AceReason-Nemotron-7B, for example, is documented as post-trained from DeepSeek-R1-Distill-Qwen-7B and is assigned to the Qwen-base-weight tier on that basis.

Table 4. Cisco MPK Nemotron reference models by base-weight tier

Tier	Model
Qwen-base-weight	nvidia/AceMath-RL-Nemotron-7B
Qwen-base-weight	nvidia/AceReason-Nemotron-1.1-7B
Qwen-base-weight	nvidia/AceReason-Nemotron-14B
Qwen-base-weight	nvidia/AceReason-Nemotron-7B
Qwen-base-weight	nvidia/Nemotron-Cascade-14B-Thinking
Qwen-base-weight	nvidia/Nemotron-Cascade-8B
Qwen-base-weight	nvidia/Nemotron-Cascade-8B-Thinking
Qwen-base-weight	nvidia/Nemotron-Orchestrator-8B
Qwen-base-weight	nvidia/Nemotron-Research-Reasoning-Qwen-1.5B
Qwen-base-weight	nvidia/Nemotron-Terminal-14B
Qwen-base-weight	nvidia/Nemotron-Terminal-8B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-1.1-14B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-1.1-7B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-14B
Qwen-base-weight	nvidia/OpenCodeReasoning-Nemotron-7B
Qwen-base-weight	nvidia/OpenMath-Nemotron-1.5B
Qwen-base-weight	nvidia/OpenMath-Nemotron-14B
Qwen-base-weight	nvidia/OpenMath-Nemotron-14B-Kaggle
Qwen-base-weight	nvidia/OpenMath-Nemotron-7B
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-1.5B
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-14B
Qwen-base-weight	nvidia/OpenReasoning-Nemotron-7B
NVIDIA-native	nvidia/Nemotron-4-Mini-Hindi-4B-Base
NVIDIA-native	nvidia/Nemotron-4-Mini-Hindi-4B-Instruct

MPK base-weight tier roster

Tier	Model
NVIDIA-native	nvidia/Nemotron-Flash-1B
NVIDIA-native	nvidia/Nemotron-Flash-3B-Instruct
NVIDIA-native	nvidia/Nemotron-Mini-4B-Instruct
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-12B-v2
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-12B-v2-Base
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-9B-v2
NVIDIA-native	nvidia/NVIDIA-Nemotron-Nano-9B-v2-Base
Llama-base-weight	nvidia/Llama-3.1-Nemotron-8B-UltraLong-1M-Instruct
Llama-base-weight	nvidia/Llama-3.1-Nemotron-8B-UltraLong-2M-Instruct
Llama-base-weight	nvidia/Llama-3.1-Nemotron-8B-UltraLong-4M-Instruct
Llama-base-weight	nvidia/Llama-3.1-Nemotron-Nano-4B-v1.1
Llama-base-weight	nvidia/Llama-3.1-Nemotron-Nano-8B-v1
Llama-base-weight	nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1

References

Aghaei, Ehsan, Amy Chang, Ankit Garg, and Sanket Mendapara. 2026. *Defining Model Provenance: A Constitution for AI Supply Chain Safety and Security*. Cisco Blogs. https://blogs.cisco.com/ai/model-provenance-constitution.

Autio, Chloe, Reva Schwartz, Jesse Dunietz, et al. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1.

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, et al. 2023. "The Foundation Model Transparency Index." *arXiv Preprint arXiv:2310.12941*, ahead of print. https://doi.org/10.48550/arXiv.2310.12941.

Bommasani, Rishi, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. 2023. "Ecosystem Graphs: The Social Footprint of Foundation Models." *arXiv Preprint arXiv:2303.15772*, ahead of print. https://doi.org/10.48550/arXiv.2303.15772.

Cheng, Pengzhou, Zongru Wu, Tianjie Ju, Wei Du, Zhuosheng Zhang, and Gongshen Liu. 2024. "Transferring Backdoors Between Large Language Models by Knowledge Distillation." *arXiv Preprint arXiv:2408.09878*, ahead of print. https://doi.org/10.48550/arXiv.2408.09878.

Cisco AI Defense. 2026a. *Model Provenance Constitution: Taxonomy, Definition, and Boundary Specification*. Model Provenance Kit documentation. https://github.com/cisco-ai-defense/model-provenance-kit/blob/main/docs/constitution/model_provenance_constitution.md.

Cisco AI Defense. 2026b. *Model Provenance Kit: Toolkit to Assess and Determine Model Provenance*. Software repository. https://github.com/cisco-ai-defense/model-provenance-kit.

European Parliament and Council of the European Union. 2024. *Regulation (EU) 2024/1689 Laying down Harmonised Rules on Artificial Intelligence*. 2024/1689. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, et al. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

Hubinger, Evan, Carson Denison, Jesse Mu, et al. 2024. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." *arXiv Preprint arXiv:2401.05566*, ahead of print. https://doi.org/10.48550/arXiv.2401.05566.

Kurita, Keita, Paul Michel, and Graham Neubig. 2020. "Weight Poisoning Attacks on Pretrained Models." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online), 2793–806. https://doi.org/10.18653/v1/2020.acl-main.249.

Meinhardt, Caroline, Sabina Nong, Graham Webster, Tatsunori Hashimoto, and Christopher D. Manning. 2025. *Beyond DeepSeek: China's Diverse Open-Weight AI Ecosystem and Its Policy Implications*. Issue Brief. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/assets/files/hai-digichina-issue-brief-beyond-deepseek-chinas-diverse-open-weight-ai-ecosystem-policy-implications.pdf.

References

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, et al. 2019. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAT* '19, 220–29. https://doi.org/10.1145/3287560.3287596.

Nguyen, Vinh, Chintan Patel, Jiaqi Zeng, and Igor Gitman. 2025. *Build Custom Reasoning Models with Advanced, Open Post-Training Datasets*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/build-custom-reasoning-models-with-advanced-open-post-training-datasets/.

NVIDIA. 2025. *Nemotron-Cascade-8B Model Card*. Hugging Face model repository. https://huggingface.co/nvidia/Nemotron-Cascade-8B.

OECD.AI. 2026. *AIKoD Data*. Methodological note for the AI Knowledge on Demand database. https://oecd.ai/en/aikod/.

Rahman, Robi. 2024. *Most Large-Scale Models Are Developed by US Companies*. Epoch AI Data Insights. https://epoch.ai/data-insights/large-scale-models-by-country.

VAIL Inc. 2026. *Technology: Behavioral Fingerprinting*. Technology documentation. https://www.projectvail.com/pages/technology.