AI SAFETY AND GOVERNANCE

Should AI Be Allowed to Use Neural State-Space Models?

A precautionary prohibition, a testable scope and a plan for independent review

Revanth Matha · Matha Labs | Preprint awaiting review 1.0 | 2026-11-20

Abstract

Should an AI service be allowed to use a learned state-space sequence layer? This paper recommends prohibiting these layers and defines the proposal so a purchaser can apply it to S4, Mamba, equivalent computations and hybrid services. A worked recurrence shows that two different observation histories can produce the same final state, leaving a reviewer unable to infer an earlier observation from that state alone. This mathematical example establishes a limit on reconstruction under specified conditions. It gives no estimate of harm across deployed architectures. We consider the prohibition's strongest rationale and its main costs, including lost efficiencies, service interruption and substitution with other systems that retain personal history. Six procurement classifications and a three-way decision procedure distinguish covered components, out-of-scope components and unresolved evidence. The proposed governance process preserves appeal, correction and continuity, and specifies how independent reviewers could compare the broad prohibition with two narrower policies.

Keywords: AI safety, state-space models, S4, Mamba, precaution, public procurement, recourse

Contributions

1. A categorical policy proposal with a technical definition covering neural sequence layers, equivalent implementations and hybrid services.

2. An exact two-history example and six illustrative classifications that distinguish eligibility, evidentiary holds and privacy requirements.

3. A comparative governance design with appeal rights, continuity safeguards and prespecified questions for revising the recommendation.

1. Plain-English summary

Imagine that a public office uses an assistant to route requests. You corrected an address this morning. By afternoon the assistant still sends your request to the wrong district. The clerk can see the corrected form, but the assistant also uses a running numerical summary of earlier interactions. You ask which earlier observation mattered and who can clear its influence. The office has to answer two questions. What did the model retain, and who gave the service permission to use it in this decision?

This paper recommends a categorical prohibition on AI systems using neural state-space sequence models. We would apply the rule to public and private services, on-device and hosted models, and components supplied by third parties. It applies even to a model that processes one document and discards its internal state at the end. The reason for proposing a broad rule is that a clear boundary can be enforced before an institution delegates decisions to a system. This breadth carries costs for useful, low-risk applications. Those costs belong in any public decision about adopting it.

Operating a state-space model and studying its mathematics are different activities. Engineers have used explicit equations to estimate the state of physical systems for decades. A well-defined instrument can continue to use such equations under the rule described here. Our subject is the use of a particular learned sequence-processing mechanism in AI. We explain the boundary, work through difficult cases, and identify evidence that could justify revising the prohibition. This is a policy recommendation for deliberation. [1][3][4]

2. How the evidence was assessed

We conducted a focused technical and policy analysis. We chose the original filtering paper to establish the older meaning of state space, the HiPPO and S4 papers for compressed sequence representations, the first Mamba paper for content-dependent selection, the duality paper for equivalent computations, and Jamba for a concrete hybrid architecture. We reviewed model reporting guidance and two voluntary NIST frameworks for organizational controls. The original Transformer provides an example of a separate model family. This is a purposive source set, selected to answer questions about mechanism and governance. [1][2][3][4][5][6][7][8][9][10]

For each source we asked what object it described, what its authors established, and what evidence a purchaser would still need. A benchmark result can support a claim about performance under a stated task. It cannot establish how many users lacked recourse in the field. A reporting framework can assign responsibilities. The framework does not itself show that an AI service complied with them. Our worked cases apply a proposed rule to stipulated facts. They are suitable for checking the rule's consistency, with no estimate of prevalence or harm inferred from their count.

Three judgments guide the recommendation. Institutions should be able to justify the history used in decisions that affect people. People should have an effective route to challenge consequential errors. The burden of demonstrating these properties should initially fall on the organizations deploying the system. A reader could accept the technical evidence while placing a different weight on these judgments. The policy comparison later in the paper makes that disagreement concrete.

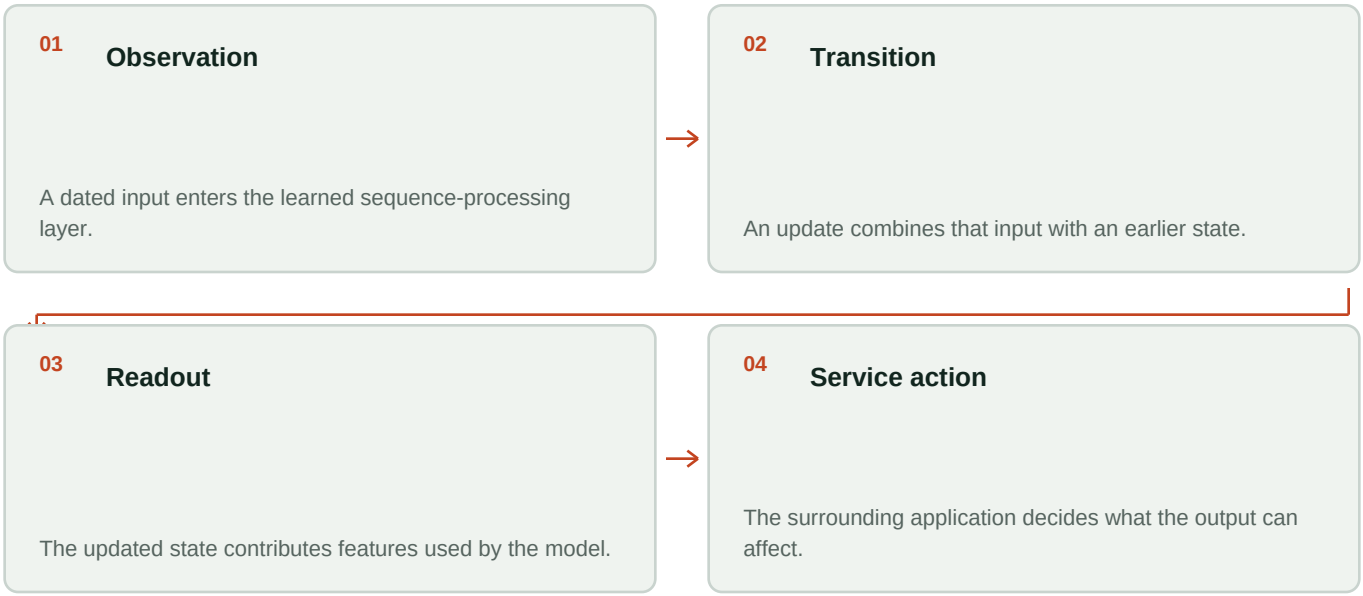
3. What counts as a neural state-space model

In a state-space description, the present state summarizes the part of the past used to compute the next state or an output. A classical estimator might maintain a position estimate and its uncertainty under a physical model. Kalman's work gave recursive equations for estimating the state of a linear system with specified assumptions about its evolution and observations. A reviewer can ask what the estimated position means, how a measurement changes it, and which assumptions make the estimator appropriate. There are serious engineering questions here even when the variables have clear physical meanings. [1]

A neural sequence model uses learned representations to process an ordered stream. In S4, a structured linear state-space layer can be evaluated as a recurrence or through an equivalent convolution for a sequence with fixed parameters. The authors develop a parameterization that makes long-range sequence modeling tractable and test it on several tasks. HiPPO gives a useful mathematical intuition for one research lineage. A finite collection of coefficients can approximate a previous signal under a chosen measure. A compressed representation can preserve information relevant to a task even when reconstructing each original observation would require additional information. [2][3]

Mamba makes selected parameters of the update depend on the current input. The contents of a stream can therefore change how information enters its state or propagates. This is a property of the computation. It does not give the model property rights, goals, a permanent user profile, or authority to take action. Whether a system carries state between sessions depends on how the service initializes, stores and reuses it. Fixed learned weights, a temporary recurrent state and a separately stored user record each have a different life cycle. Governance has to identify all three. [3][4][9]

Figure 1. Where sequence history can influence a decision



The drawing follows one inference path. A persistent store, a session boundary and an authority to act each require their own description. A sequential dependency can exist even when the implementation computes the sequence in parallel. [3][4][5]

4. The proposed rule and its boundary

An adopter should prohibit an AI system from invoking, embedding, or depending on a neural state-space sequence layer for inference or automated action. For this rule, a covered layer is a sequence-processing component in a learned model that carries information through a structured linear or input-conditioned linear state transition over ordered observations and maps that trajectory to learned features or outputs. Coverage includes an equivalent convolution or structured operator that implements the covered layer without materializing each intermediate state. Nonlinearity elsewhere in the model leaves the status of that component unchanged. The definition targets the S4 and selective-SSM lineage. [3][4][5]

The system boundary extends to a subcontractor whose covered output enters the AI service. A supplier may use a different name for the product or advertise an attention-based front end. The relevant question is what the active version actually executes. In Jamba, for example, attention and Mamba layers appear in the same model. An adopter of the proposed rule would exclude that configuration. A purchaser that cannot classify a component must hold authorization while the evidence is missing. Such a hold records uncertainty and can be lifted when the supplier substantiates the architecture. [5][6]

A fixed physical estimator whose equations and state variables are prescribed independently of neural sequence learning falls outside this definition. A downstream AI model may consume a documented measurement from that instrument. An AI system that inserts a learned selective SSM into the measurement path falls within the prohibition. A cached attention layer requires review of its own retention behavior. Mathematical resemblance at a high level, or the mere use of finite memory, would be an unsound

classification test. The duality work connects particular classes of structured operators and attention variants. A reviewer must identify the realized operator and training role. [1][5][10]

Table 1. Eligibility under the proposed architectural rule

Stipulated component	Architectural outcome	Reason or next question
Fixed physical estimator with documented equations	Outside scope	Physical dynamics and an explicit measurement model prescribe the update. [1]
S4 encoder evaluated as a convolution	Covered	The active operator implements the covered structured sequence layer. [3]
One-session Mamba model with reset at the boundary	Covered	The reset addresses persistence; the learned selective update remains active. [4]
Attention-Mamba hybrid model	Covered	A covered layer contributes to the active inference path. [6]
Conventional attention with a separately retained user database	Outside this gate	Review the data store, actions and recourse under other controls. [9][10]
Hosted summarizer with insufficient component evidence	Hold	The purchaser needs evidence about the shipped version and its dependencies. [7][8]

These are worked classifications of specified configurations, not findings about six deployed products. Outside scope signifies eligibility for further review. It conveys no safety certification.

5. What a running state can and cannot establish

Consider a transparent teaching example. A scalar state starts at zero. At each step the new state equals one half of the old state plus the new observation. The histories (1, 0) and (0, 0.5) both end at 0.5 after two observations. A decision rule using only the final state must give them the same answer. Yet a reviewer asking whether the first observation was one would need to distinguish them. The ambiguity follows exactly from the stipulated update. The example is a many-to-one mapping, built to isolate this property. It supplies no frequency estimate for ambiguous decisions made by deployed neural models.

Now imagine correcting the first history by replacing its initial one with zero. Under this example's equations, replaying the corrected trace gives a final state of zero. A service that stored only the final 0.5 would need more evidence to establish which history it should replay. Keeping all raw inputs could enable that particular reconstruction and create a new record of a person's behavior. A sound correction policy must say which facts are retained, who can authorize replay, who may view the evidence and when the material expires. The right balance depends on the purpose and the consequences for the person. [9]

This example distinguishes auditability from prediction quality. If the assigned prediction depends only on the final scalar, the model might answer every test question correctly while failing every attempt to recover the first observation from that scalar alone. A task which scores final actions will overlook the missing evidence.

Conversely, an independently logged observation may establish why a rule applied without requiring anyone to interpret a latent coordinate. The deployment design therefore matters to both the case for a ban and the case for a narrower alternative. [2][8]

```
state(t) = 0.5 x state(t-1) + observation(t); state(0) = 0
```

For the sequences (1, 0) and (0, 0.5), the final state equals 0.5 in both cases. Inferring the first observation from that one scalar has no unique solution.

6. Why propose a categorical prohibition

A broad prohibition has a specific institutional rationale. A purchaser can require an architectural determination before signing a contract, while a person harmed by a decision may be unable to reconstruct the service that made it. The adopter moves the burden of discovery upstream to a party able to inspect the design. The policy also limits dependence on a vendor's claim that hidden sequence state will serve only harmless purposes after an integration changes. It makes the same covered component ineligible when purchased outright, embedded in a hybrid, or reached through another service.

The performance studies supply a reason to take the deployment question seriously. Efficient long-sequence processing can make use of longer or more frequent observation streams feasible under a given computational budget. That is a capability claim conditional on an implementation and workload. It does not show that an operator will collect additional personal data, that individuals will suffer a particular loss, or that useful efficiencies should always be sacrificed. The policy case rests on the burden of assuring oversight as deployments spread, combined with the value we assign to meaningful recourse. [3][4][8][9]

The breadth is deliberate. An operator could process public environmental data with a covered model, discard all state between requests and expose a plain numerical forecast. The proposed prohibition would still reach that service. We accept this cost in the proposal because an architecture-wide rule is easier for purchasers to state consistently and gives a regulator a clear choice about changing the category. Whether it is easier to enforce in practice remains an empirical question. An inaccessible hosted implementation can defeat architectural inspection just as it can frustrate an audit of state retention.

7. Applying the rule to procurement decisions

Suppose a school buys a translation service for public notices. The vendor documents a Mamba encoder and guarantees that every request begins with a fresh state. The reset guarantee is relevant to privacy. The covered encoder still requires the school to decline that configuration under the rule. The school should first identify the communication obligation, the languages required and the greatest tolerable delay. A replacement qualifies only after staff assess translation quality and whether people without connectivity can still obtain the notices. Changing the architecture alone does not complete the service transition. [4][7]

Consider a flood monitoring station with an explicitly programmed estimator that reduces fluctuations in gauge readings. An independent forecasting application receives a water-level estimate and its timestamp. The equipment record identifies the physical state variables, the update equations and the input interface. Under the stipulated facts the estimator is outside the proposed ban. If a vendor update quietly replaces that stage with a

learned S4 block, the changed component requires a fresh determination. The reviewing office must preserve the original safety obligations for the flood service throughout the change. [1][3]

A third purchaser uses a general assistant and receives a model card with impressive language scores, a version name and intended uses. The vendor cannot identify the architecture behind a summarization subcontractor. Here the correct immediate outcome is to hold authorization pending documentation. If the subcontractor later attests, with reviewable evidence, that the active configuration uses a covered hybrid, the hold becomes a denial. If it supplies evidence of an ordinary attention model, the architecture question can close and the purchaser proceeds to an ordinary assessment of what historical information the service uses. [6][7][8]

8. The strongest case for a narrower rule

An architecture-neutral alternative would require every consequential AI service to disclose its data-retention practices, separate users' sessions, limit access to stored history, and provide an appeal and correction route. Its central advantage is coverage of functionally similar outcomes. An attention model connected to a long-lived record of users may shape later decisions about them. Replacing a covered model with that system could preserve the very problem the ban aims to address. A smaller covered model run locally could use fewer resources or require less disclosure to remote vendors than its replacement. Those possibilities merit testing for the particular service. [3][4][9][10]

A deployment-specific restriction is a second alternative. An adopter could restrict AI use of covered models in consequential decisions or when personal history persists across sessions. That would retain more latitude for environmental monitoring or isolated analyses. Its weakness is the boundary between an advisory score and a real decision. A staff member may nominally approve every case while relying heavily on an automated ranking. A rule defined by consequences needs evidence about how staff actually use the output, including the remedy offered to a person whose case is misclassified. [8]

The categorical proposal should stand or fall on a comparison with these alternatives. It is strongest where architectural evidence is available before procurement, harmful downstream history use is hard to supervise, and an accessible replacement can deliver the service. It is weakest where the dominant harm arises from an external database or delegation of authority independent of the model family. A hybrid policy can impose the categorical gate and retain architecture-neutral privacy and action controls. Passing the gate would settle only the narrow eligibility question. [5][8][9]

Table 2. Comparing ways to regulate historical influence

Policy	Evidence it needs	Main risk to evaluate
Prohibit covered neural SSM components	A determination of what the active system implements	Substitution may preserve harmful history use and reduce service quality.
Restrict specified deployments	Evidence about session boundaries and how outputs affect people	Changing use or nominal human approval can obscure the decision boundary.
Require recourse and controls across architectures	Verified retention, correction, review and remedy throughout the service	Enforcement may occur only after a person has already suffered an error.

These alternatives reward different kinds of evidence. Comparing their effects requires observing services and affected people. This is a proposed evaluation design. [7][8][9]

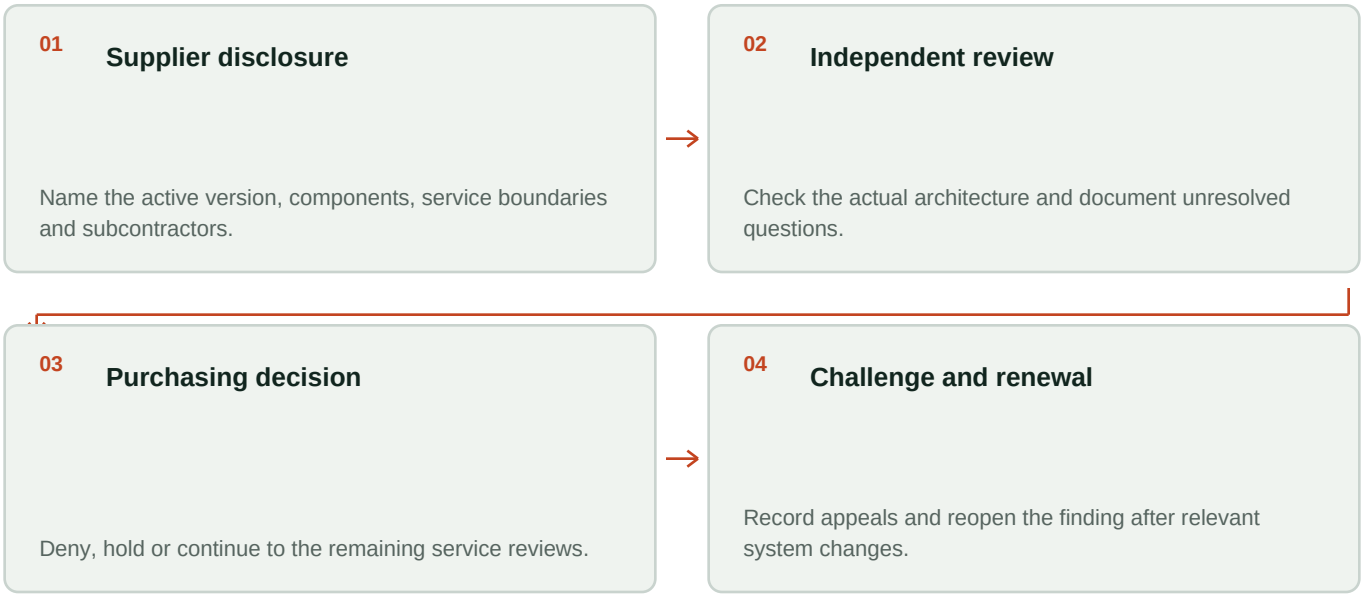
9. Who decides, who can challenge, and what gets recorded

Under the proposed adoption procedure, a supplier identifies each version and component in the stream-processing path, the role of the component in inference, and any external service that transforms inputs. A technically qualified reviewer compares the declaration with an independently inspectable artifact or a confidential third-party assessment. Trade secrecy can change who inspects the design. It should leave the purchaser with a signed, appealable finding, its scope, and the evidence needed to revisit it. A model card is a useful place to identify intended use and limitations, alongside this separate determination of architectural eligibility. [7]

The purchasing authority then records a three-way disposition. A covered component leads to denial. Conclusive evidence of an out-of-scope component allows other service reviews to continue. Insufficient or contradictory evidence leads to a hold with a stated question and a named reviewer. The decision record should specify the configuration, date, authority, reasoning, and triggers for renewal. A supplier may challenge a mistaken classification by providing new technical evidence. A person affected by the resulting service also needs a distinct channel for challenging its decisions. The purchaser owns continuity and correction even when it delegates technical inspection. [7][8]

Existing deployments require an orderly transition. An adopting authority would stop new uses and expansion, inventory existing configurations, identify the service owner, and set a withdrawal plan with named milestones. For an essential service, a human or independently validated replacement should be available before a cutover that could interrupt service. A narrowly time-limited transition, publicly justified by the adopting authority, addresses continuity for an installed system. It gives a vendor no continuing entitlement to sell the covered configuration or extend the deployment. People who depend on the service should have a direct route to report degradation during the transition. [8]

Figure 2. A proposed purchasing and review procedure



This proposed allocation of responsibilities draws on model reporting and risk management. The cited authors and NIST do not thereby endorse a categorical prohibition. [7][8]

10. Evidence that could change the decision

A public evaluation should compare three policies on the same set of service workflows. One adopts the categorical prohibition with service continuity measures. Another uses a restriction tied to persistent state or consequential decisions. The third applies architecture-neutral requirements for recourse, evidence and retention. Record the decision context, budgets, population served, action authority and revision history before comparing outcomes. Where possible, reviewers should examine existing deployments, consenting test settings and archived artifacts. A jurisdiction that adopts the prohibition can analyze previously produced results and permitted comparators without itself operating a new prohibited service.

Measure recourse in terms that a resident could recognize. Of all contested decisions, how often did a reviewer obtain sufficient evidence to explain the disposition? Among confirmed errors, what fraction received timely correction? Among requests to clear contextual history, what proportion showed the expected effect in a consented repeat evaluation, with each independent store accounted for? Publish denominators, delays, unavailable evidence, false holds, and service interruption. Report error and delay by service and affected group where lawful, consented and statistically supportable. Evaluation records should use only the information needed for their purpose. [7][8][9]

A fair review also measures human costs of substitution. Record energy under a common hardware and workload protocol, maintenance labor, service delay, and the additional personal information collected by any replacement. An independent panel should revisit the rule on a declared schedule and publish its reasons for continuation, amendment or repeal. Evidence that narrower rules provide equivalent recourse with lower

service losses would weaken the categorical case. Repeated inability to determine what influenced consequential decisions despite targeted safeguards would strengthen it. Neither outcome can be inferred from language benchmark scores alone. [3][4][8]

11. Limits and conclusion

No field study in this paper measures whether neural SSMs increase harmful decisions relative to attention models, databases, or manually administered policies. The derivation shows a specific loss of recoverable history under a simple recurrence. The procurement cases show how a proposed definition behaves under stated facts. Changes in model design may introduce operators that are hard to classify under that definition. Independent adjudication and publication of borderline cases would be necessary for consistent enforcement.

Our recommendation remains direct. An adopting institution should prohibit AI systems from using learned neural state-space sequence layers, including selective and hybrid versions. It should distinguish that prohibition from oversight of conventional physical estimators, fund continuity for the people who rely on existing services, and enforce ordinary privacy and action controls on allowed replacements. The public should be able to ask whether those steps improved the ability to challenge a decision. That answer requires evidence from people and institutions, as well as knowledge of the computation inside the model.

References

[1] Kalman RE. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, 1960. Primary mathematical account of recursive estimation for linear dynamical systems. Accessed 2026-11-20.

[2] Gu A, Dao T, Ermon S, Rudra A, Ré C. HiPPO: Recurrent Memory with Optimal Polynomial Projections. NeurIPS 2020. Defines projection of a history onto a finite polynomial basis under a specified weighting measure. Accessed 2026-11-20.

[3] Gu A, Goel K, Ré C. Efficiently Modeling Long Sequences with Structured State Spaces. ICLR 2022. Describes S4, its recurrent and convolutional views, and long-sequence benchmark results. Accessed 2026-11-20.

[4] Gu A, Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Original Mamba paper. Introduces input-dependent selection and evaluates sequence modeling and implementation efficiency. Accessed 2026-11-20.

[5] Dao T, Gu A. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. ICML 2024. Establishes mathematical connections for structured operators and introduces Mamba-2. Accessed 2026-11-20.

[6] Lieber O and colleagues. Jamba: A Hybrid Transformer-Mamba Language Model. Primary architecture report, 2024. Describes interleaving attention and Mamba layers in a single model. Accessed 2026-11-20.

[7] Mitchell M and colleagues. Model Cards for Model Reporting. FAT* 2019. Proposes reporting intended uses, performance conditions and limitations for trained models. Accessed 2026-11-20.

[8] Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023. Voluntary framework organizing risk work into governance, context mapping, measurement and management. Accessed 2026-11-20.

[9] NIST. Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management, Version 1.0. January 2020. Voluntary framework for identifying data-processing risks to individuals and managing them across a service. Accessed 2026-11-20.

[10] Vaswani A and colleagues. Attention Is All You Need. NeurIPS 2017. Presents the original Transformer design and empirical translation evaluation. Accessed 2026-11-20.

Reproducibility record

Private review materials: https://github.com/RevanthM/ssm-safety | Research snapshot: 2fa622e435946d0f1aa77813c82500ded3595c09