

Sufficient Projections

Flyxion
Independent Researcher

October 30, 2026

Abstract

A system that compresses, summarizes, or reformulates a record is usually judged by how well the reduced representation serves the questions asked of it at the time of reduction. This essay treats a different requirement: that the reduced representation remain adequate for judgments that have not yet been specified, namely those generated when the governing policy is later revised. I define sufficiency of a projection relative to a class of admissible revisions, show that it is characterized exactly by a verdict-equivalence relation on histories, and derive three consequences: the minimal cost of sufficiency is the number of verdict classes the revision class induces, a test collection certifies sufficiency for the class it was drawn from exactly when it induces the same verdict-equivalence relation, so that the operative condition is the relation and not the size of the collection, and the failure modes of retention, formalization, and execution are logically independent. Four recent results, on semantic alignment of formal statements, thread-structured agent memory, layered misalignment monitoring, and agent-written infrastructure, are read as instances of this structure. A new policy can be admitted to the family a sufficient projection supports exactly when the projection's kernel lies within the policy's kernel.

1 The Problem

Every system that acts on a record faces a choice about what to keep. The full record is expensive to store, expensive to search, and often too large to place in front of whatever process must decide something at the present moment. A summary is cheap. The ordinary standard for a summary is local: it is good if the decisions made from it agree with the decisions that would have been made from the full record, for the questions that were in view when the summary was constructed. This standard is entirely reasonable and entirely insufficient for any system whose governing rules can change, because a rule that did not exist when the summary was made may depend on a distinction the summary discarded.

The difficulty is not that summaries lose information. That is what summaries are for. The difficulty is that what is lost is selected by the policy in force at the time of loss. The loss is therefore invisible to any evaluation conducted under that same policy. A reduction that passes all present checks has been tuned to those checks. When a revision arrives requiring something that was peripheral under the old policy, the reduced representation may be unable to supply it, and the system cannot discover this by inspecting the representation, since what is absent leaves no trace to inspect.

This essay isolates that situation and states it precisely. The aim is structural. I do not propose a memory architecture, a monitoring scheme, or a verification tool. I define when a

projection of a history is sufficient for a class of future revisions, characterize that property exactly, and then use the characterization to read several recent research results as instances of a single phenomenon.

2 Antecedents in Statistics, Automata, and Databases

The central notion is a descendant of several older ones, and the formal results below are elementary consequences of them. In estimation theory a statistic $T(X)$ is sufficient for a parameter when it retains everything the sample X says about that parameter [6, 7], and a minimal sufficient statistic is the coarsest such reduction [8]. The correspondence with the present setting is close: the factorization of a likelihood through T is replaced by the factorization of a policy through a projection, and the minimal sufficient statistic is replaced by the quotient by the verdict-equivalence defined in Section 3. Blackwell’s comparison of experiments [9] supplies the ordering view, in which one reduction is at least as informative as another when the second factors through the first. The monotonicity in Corollary 1 is a cousin of this ordering: enlarging the revision class enlarges the set of policies that must factor through the projection.

In automata theory the Myhill and Nerode congruence characterizes the minimal automaton for a language as the quotient of histories by the relation that agrees on every possible continuation [10]. The relation $\sim_{\mathcal{R}}$ quantifies over policies where the congruence quantifies over continuations. Both identify the coarsest equivalence that preserves their respective observations, but $\sim_{\mathcal{R}}$ need not be compatible with concatenation of histories, which is the defining property of a right congruence. The two coincide when the policy family induces precisely the continuation-indexed distinctions relevant to the language, and otherwise the constructions are analogous and not identical. Under that qualification Proposition 1 is the corresponding statement that a reduction is adequate exactly when it separates no pair the congruence separates. The closest database analogue is view determinacy, where a set of views determines a query when the query can be computed from the view images alone [11]. Sufficiency in the sense used here is determinacy of a whole class of queries by a single view, and the proposition on certification asks when a sample of queries suffices to decide determinacy for the class. In machine learning the information bottleneck [12] treats compression relative to a relevance variable, and the present account differs by requiring exact preservation of verdicts for every member of a class instead of a trade-off between compression and relevance. Abstraction refinement in verification [13] addresses the same tension from the other direction, by refining an abstraction when a spurious counterexample shows that it merged histories a property distinguishes.

None of the formal results is claimed as new mathematics. What the essay adds is the relativization to an admissible revision class that the governing authority may extend, the consequence that the distinctions a reduction must retain cannot be determined from the present policy alone, since they depend on the range of policies the authority may adopt as well as on the histories themselves, and the separation of historical, semantic, and operational adequacy as independent requirements. The applied works discussed in Section 8 are not antecedents of the theory. They are read as candidate instances of the pattern, and the claim made about them is structural resemblance, not empirical support.

3 Definitions

Let $\mathcal{H}$ be a set of histories, understood as complete records of what occurred, with no commitment yet to their internal structure. Let $\mathbb{V}$ be a set of verdicts, for instance permitted, deferred, or refused. A *policy* is a function $\rho : \mathcal{H} \rightarrow \mathbb{V}$ assigning a verdict to each history. Let $\mathcal{R}$ be a set of policies, the *admissible revisions*: those the system's governing authority could in principle adopt. The authority is whatever body sets policy for the system; its range of admissible action is the object of interest, not any particular policy it currently holds. The present policy is one member of $\mathcal{R}$, but nothing below requires it to be distinguished.

A *projection* is a function $\pi : \mathcal{H} \rightarrow P$ into some set P of reduced representations. A projection may replace the history in storage, or it may be a temporary working representation while the authoritative history is retained separately. A policy ρ is *determined by* π if there is a function $\hat{\rho} : P \rightarrow \mathbb{V}$ with $\hat{\rho} \circ \pi = \rho$, so that the verdict is determined by the reduced representation alone, in the sense that it factors through π . Factoring asserts the existence of a function and does not assert that the function is effectively computable.

Definition 1 (Sufficiency). *A projection π is sufficient for $\mathcal{R}$ if every $\rho \in \mathcal{R}$ is determined by π .*

Define the *verdict-equivalence* of $\mathcal{R}$ on $\mathcal{H}$ by

$$h \sim_{\mathcal{R}} h' \iff \rho(h) = \rho(h') \text{ for all } \rho \in \mathcal{R}.$$

This relation records which distinctions between histories could ever matter to a judgment the authority is able to make. Two histories related by it are interchangeable for every admissible policy. Two histories not related by it differ in some way that at least one admissible revision would act upon.

4 The Characterization

Write $\ker \pi$ for the equivalence relation $h \ker \pi h' \iff \pi(h) = \pi(h')$, the pairs of histories the projection fails to tell apart.

Proposition 1. *A projection π is sufficient for $\mathcal{R}$ if and only if $\ker \pi \subseteq \sim_{\mathcal{R}}$.*

Proof. Suppose π is sufficient and $\pi(h) = \pi(h')$. For each $\rho \in \mathcal{R}$ there is $\hat{\rho}$ with $\rho = \hat{\rho} \circ \pi$, hence $\rho(h) = \hat{\rho}(\pi(h)) = \hat{\rho}(\pi(h')) = \rho(h')$. Therefore $h \sim_{\mathcal{R}} h'$. Conversely suppose $\ker \pi \subseteq \sim_{\mathcal{R}}$ and fix $\rho \in \mathcal{R}$. Define $\hat{\rho}$ on the image of π by $\hat{\rho}(\pi(h)) = \rho(h)$. This is well defined, since $\pi(h) = \pi(h')$ implies $h \sim_{\mathcal{R}} h'$ and so $\rho(h) = \rho(h')$. Extend $\hat{\rho}$ to all of P by assigning arbitrary verdicts outside $\pi(\mathcal{H})$; this requires only that $\mathbb{V}$ be nonempty. Then $\hat{\rho} \circ \pi = \rho$. $\square$

The proposition says that a projection is sufficient exactly when it never merges two histories that some admissible revision would treat differently. The condition is a refinement condition, not a similarity condition. It is not enough that the projection preserve what is typically relevant, or what is relevant to likely revisions. Nor is it enough that it preserve what a retrieval procedure would rank highly for a representative query. A single admissible policy that separates two merged histories defeats sufficiency, however improbable adoption of that policy may be.

Corollary 1 (Cost). *If π is sufficient for $\mathcal{R}$, then $|\pi(\mathcal{H})| \geq |\mathcal{H}/\sim_{\mathcal{R}}|$, and the bound is attained by the quotient map. Moreover, if $\mathcal{R} \subseteq \mathcal{R}'$ then $\sim_{\mathcal{R}'} \subseteq \sim_{\mathcal{R}}$, so the minimal sufficient representation for $\mathcal{R}'$ is at least as large as for $\mathcal{R}$.*

Proof. By Proposition 1, $\ker \pi$ refines $\sim_{\mathcal{R}}$, so π has at least as many fibres as $\sim_{\mathcal{R}}$ has classes. The quotient map $h \mapsto [h]_{\mathcal{R}}$ has kernel exactly $\sim_{\mathcal{R}}$ and is therefore sufficient. If $\mathcal{R} \subseteq \mathcal{R}'$, agreement on all of $\mathcal{R}'$ implies agreement on all of $\mathcal{R}$. $\square$

When $\mathcal{H}/\sim_{\mathcal{R}}$ is finite, a fixed-length encoding of its classes needs at least $\lceil \log_2 |\mathcal{H}/\sim_{\mathcal{R}}| \rceil$ bits in the worst case. The corollary itself bounds the cardinality of the image and says nothing directly about storage, computational, or retrieval cost, which depend on the encoding and on how the classes are accessed. The corollary gives the cost of future admissibility a definite form. Enlarging the class of revisions the authority is permitted to make can only enlarge the number of distinctions that must survive. The cardinality $|\mathcal{H}/\sim_{\mathcal{R}}|$ depends on both the domain of histories and the policy family, so the cost is not a property of the present policy and cannot be read off from it. What the present policy cannot reveal is the set of distinctions that other admissible policies would use, and that set is a feature of the authority’s range of future action.

Remark 1. A system may reduce this cost by constraining the revision class, for instance by requiring that new policies be expressible over a fixed vocabulary of retained features. This is a legitimate design, but it moves the burden rather than removing it: the vocabulary becomes the object whose adequacy is in question, and a revision that cannot be expressed in it is excluded by construction, not accommodated.

Remark 2 (The verdict set). The set $\mathbb{V}$ is part of the specification, not a fixed background. Yang, Chen, and Gao describe OA-RAG, which grounds retrieval-augmented answers in formal ontologies and distinguishes conclusions entailed by the ontology from conclusions merely supported by related evidence [5]. Distinguishing entailment from association enlarges $\mathbb{V}$, so that a history can warrant a conclusion in more than one way. Enlarging $\mathbb{V}$ while holding the policy family fixed changes no equivalence class, since the policies assign the same verdicts as before. A refinement of $\sim_{\mathcal{R}}$ occurs only when an admissible policy uses the enlarged verdict set to separate histories that were previously treated alike, and a projection tuned to the coarser verdict set may then prove insufficient. Part of the work of preserving future admissibility therefore lies in the verdict set and its use by policies, and not only in the projection. The reported grounding categories include model-based judgments, so the arrangement mixes formal and informal warrant and should be read as such.

5 Test Collections and Their Classes

In practice sufficiency is rarely checked against $\mathcal{R}$ itself, since $\mathcal{R}$ may not be enumerable. It is checked against a smaller collection $T \subseteq \mathcal{R}$ of policies, test cases, or benchmark questions. Sufficiency for T means $\ker \pi \subseteq \sim_T$, and since $T \subseteq \mathcal{R}$ the relation $\sim_T$ contains $\sim_{\mathcal{R}}$. Whether it is strictly coarser is the whole question.

Proposition 2. *If $T \subseteq \mathcal{R}$ and $\sim_T \neq \sim_{\mathcal{R}}$, then there is a projection that is sufficient for T and not for $\mathcal{R}$.*

Proof. Take π to be the quotient map for $\sim_T$. Its kernel is $\sim_T$, which is contained in itself, so π is sufficient for T . Since $\sim_T \neq \sim_{\mathcal{R}}$ and $\sim_{\mathcal{R}} \subseteq \sim_T$, we have $\sim_T \not\subseteq \sim_{\mathcal{R}}$, so by Proposition 1 the projection is not sufficient for $\mathcal{R}$. $\square$

The proposition establishes existence and no more. It does not show that every procedure optimizing against T converges on such a projection, and it does not show that a collection smaller than $\mathcal{R}$ can never certify sufficiency. Finiteness is not the operative condition. For a concrete instance, let $\mathcal{H} = \mathbb{V} = \mathbb{N}$ and let $\mathcal{R}$ be any infinite family of policies that contains the identity map. The singleton $T = \{\text{id}\}$ already induces the same relation as $\mathcal{R}$, since the identity separates every pair of histories, and a projection sufficient for T is then sufficient for the whole family. The operative condition is whether $\sim_T = \sim_{\mathcal{R}}$.

Proposition 3 (Certification). *Let $T \subseteq \mathcal{R}$. Every projection sufficient for T is sufficient for $\mathcal{R}$ if and only if $\sim_T = \sim_{\mathcal{R}}$.*

Proof. If $\sim_T = \sim_{\mathcal{R}}$, Proposition 1 gives the claim directly. If $\sim_T \neq \sim_{\mathcal{R}}$, Proposition 2 supplies a projection sufficient for T and not for $\mathcal{R}$. $\square$

The forward direction is the substantive one: if $\sim_T \neq \sim_{\mathcal{R}}$, Proposition 2 exhibits a projection that is sufficient for T and merges histories that $\mathcal{R}$ distinguishes. A test collection is therefore adequate precisely when it preserves the distinctions induced by the entire admissible revision class. The obstacle in practice is the absence of a separating basis for $\mathcal{R}$, not finiteness as such.

6 Admitting a Revision

The characterization turns revision into an explicit event with an exact test. Suppose a projection is already sufficient for a policy family and the authority proposes a new policy.

Proposition 4 (Admission). *Let π be sufficient for $\mathcal{R}$ and let ρ be any policy. Then π is sufficient for $\mathcal{R} \cup \{\rho\}$ if and only if $\ker \pi \subseteq \ker \rho$.*

Proof. Agreement on every member of $\mathcal{R} \cup \{\rho\}$ is agreement on every member of $\mathcal{R}$ together with agreement under ρ , so $\sim_{\mathcal{R} \cup \{\rho\}} = \sim_{\mathcal{R}} \cap \ker \rho$. By Proposition 1, π is sufficient for $\mathcal{R} \cup \{\rho\}$ exactly when $\ker \pi \subseteq \sim_{\mathcal{R}} \cap \ker \rho$. Since $\ker \pi \subseteq \sim_{\mathcal{R}}$ by hypothesis, this holds exactly when $\ker \pi \subseteq \ker \rho$. $\square$

The proposition makes policy revision an admission event instead of a reconstruction event. A revision need not trigger a rebuild of the history or of the projection. It triggers a factorization obligation: if the obligation is met, the projection remains authorized for the enlarged family, and if it fails, the revision must be evaluated from a more informative representation. Failure does not imply that the full history is required, since another retained projection may already be sufficient. Ordering representations by informativeness, so that π' is at least as informative as π when $\ker \pi' \subseteq \ker \pi$, gives a hierarchy in which admission selects the least informative representation whose sufficiency for the revised family can be established.

The price of a revision can be stated exactly when the quotient is finite. Write $N(\mathcal{R}) = |\mathcal{H}/\sim_{\mathcal{R}}|$. Adding a policy refines the existing partition, and if k_C denotes the number of distinct verdicts that ρ assigns to the histories in an existing class C , then $N(\mathcal{R} \cup \{\rho\}) = \sum_C k_C$, the sum running over the classes of $\sim_{\mathcal{R}}$. A new policy is free in representational terms precisely when

every $k_C = 1$, and otherwise it splits classes. If four classes become one class split three ways, another split two ways, and two classes left intact, the minimum number of representations rises from four to $3 + 2 + 1 + 1 = 7$, and a fixed-length encoding rises from two bits to three. The quantity $\log_2(N(\mathcal{R} \cup \{\rho\})/N(\mathcal{R}))$ measures the growth in the logarithm of the minimum number of states and is not the growth in encoded storage, since fixed-length encodings involve ceilings.

The cost of a policy is therefore contextual. For two families whose joint partition is the common refinement of their individual partitions, $\max(N_1, N_2) \leq N_{12} \leq N_1 N_2$, and the upper bound is attained when every combination of classes is realized by some history. Histories with two independent binary attributes illustrate this: a family distinguishing only the first attribute and a family distinguishing only the second each require two classes, and together they require four, which attains the upper bound. A second family that again distinguished only the first attribute would give $N_{12} = 2 = \max(N_1, N_2)$, attaining the lower bound. An institution can accumulate a large retention obligation through many individually modest additions, while a policy that distinguishes nothing the institution does not already distinguish costs nothing.

The final partition is independent of the order in which policies are added, since intersection is commutative and associative, but reachability is not. If an institution discards its history after compressing it by π_0 , a later projection π_1 can be constructed from π_0 exactly when the following holds.

Proposition 5 (Reachability). *Let $\pi_0 : \mathcal{H} \rightarrow P_0$ and $\pi_1 : \mathcal{H} \rightarrow P_1$ with P_1 nonempty. There is a function $f : P_0 \rightarrow P_1$ with $\pi_1 = f \circ \pi_0$ if and only if $\ker \pi_0 \subseteq \ker \pi_1$.*

Proof. If $\pi_1 = f \circ \pi_0$ and $\pi_0(h) = \pi_0(h')$, then $\pi_1(h) = f(\pi_0(h)) = f(\pi_0(h')) = \pi_1(h')$. Conversely, if the inclusion holds, define $f(\pi_0(h)) = \pi_1(h)$, which is well defined because $\pi_0(h) = \pi_0(h')$ implies $\pi_1(h) = \pi_1(h')$, and extend f arbitrarily outside the image of π_0 . $\square$

The two factorization results answer different questions. Proposition 1 asks whether a policy can be evaluated from a projection, which is decision sufficiency, and Proposition 5 asks whether one representation can be constructed from another, which is reconstruction sufficiency. An authoritative history supports the construction of any new projection, while an old projection may be unable to reconstruct a revised one. When the inclusion fails, the sufficient representation exists mathematically and cannot be constructed from what is retained. Two institutions with the same current summary and the same current policies may therefore differ completely in future capability, according to whether one of them kept an authoritative history.

Two senses of admissibility should also be separated. Let $\mathcal{R}_A$ be the policies the authority is authorized to adopt, and let $\mathcal{R}_\pi = \{\rho : \ker \pi \subseteq \ker \rho\}$ be the policies a given projection supports. The policies that are both authorized and evaluable are exactly $\mathcal{R}_A \cap \mathcal{R}_\pi$. An authority can hold the formal power to revise its rules while lacking the information needed to apply them to past events, and a projection can support policies that the institution has no authority to adopt. Formal authority does not imply informational capacity, and informational capacity does not imply authorization.

Two limiting cases show how little present performance reveals. If the family consists, for every history h , of the policy that returns one for h and zero elsewhere, every verdict is a single bit and yet $\sim_{\mathcal{R}}$ is the identity, so every sufficient projection is injective; the simplicity of individual verdicts does not bound the record needed to support them collectively. At the other

extreme, a current policy that assigns one verdict to every history is satisfied perfectly by a constant projection, which achieves maximal compression and becomes insufficient the moment any later policy separates two histories. A system can satisfy its entire present specification, with no failed test and no loss of measured performance, while losing the capacity to satisfy a future one.

7 Three Adequacies

Sufficiency as defined concerns whether a representation supports verdicts. It sits within a larger chain in which three separate things must hold, and the chain is only as strong as the weakest of them. First, the relevant evidence must have been retained, which I call historical adequacy. Second, the representation in which a constraint is written must express the distinction the constraint was meant to enforce, which I call semantic adequacy. Third, execution must in fact be governed by the constraint as represented, which I call operational adequacy.

These are logically independent. A complete history can be read under a representation that omits the distinction that matters, so historical adequacy holds while semantic adequacy fails. A perfectly faithful formalization can be evaluated against a projection that has already discarded the evidence it needs, so semantic adequacy holds while historical adequacy fails. A correctly formalized constraint evaluated against complete evidence can fail to bind the action that follows, so both of the first two hold while operational adequacy fails. None of the three implies another: a guarantee of one is not a guarantee of the others, and a failure in any one yields an unjustified judgment regardless of the other two. The characterization of Section 4 addresses the first. The following section shows that recent work addresses each of the three, and in each case an assurance obtained in one respect is at risk of being mistaken for an assurance in another.

8 Four Recent Instances

The items discussed here are taken from a research shortlist dated October 30, 2026. I describe them as their authors are reported to describe them, and I use them as illustrations of the structure above, not as evidence for it. The structure is established by the preceding propositions, which do not depend on any empirical result.

8.1 Semantic Adequacy of Formal Statements

Shen, Qiu, Zhi, and Gao introduce MechAlign, a procedure for evaluating whether a formalized mathematical result expresses the claim its authors intended, by reconstructing the dependency structure of definitions and declarations on which a Lean statement rests and comparing it with the source manuscript [1]. The work is reported to examine 201 main results from 173 projects and to assemble a benchmark of 501 claim, statement, and context triples. One reported case concerns a formal statement lacking a dimensional constraint present in the manuscript, in a project whose declared scope acknowledges the omission. The case is not offered as evidence that formalization is unreliable; it illustrates a semantic distinction that survived in the manuscript and did not survive translation.

The relevance is by analogy, and the analogy can be made literal only under stated conditions. To apply Proposition 1 one needs a domain of informal claims serving as $\mathcal{H}$, a map from claims to formal statements serving as π , and a family of judgments on claims serving as $\mathcal{R}$. Where those are supplied, a formal statement that merges two informal claims differing in a missing dimensional constraint is insufficient for any family containing a judgment that separates them, and no amount of mechanical validity repairs this. Without them, the case is best read as an instance of semantic distinctions lost between a claim and its formalization. In either reading a proof checker establishes that a proposition follows from its hypotheses, and cannot establish that the proposition is the one intended, since that is a relation between the formal statement and something outside the formal system. The failure is one of semantic adequacy.

8.2 Historical Adequacy and Thread Structure

Zhang and colleagues propose TrawMem, which organizes interaction histories into persistent threads and assembles task-specific workspaces of complementary memories instead of retrieving by similarity to the current query [2]. Macro-F1 improvements over representative alternatives are reported as 2.23 to 3.14 points. Some configurations use fewer tokens, though the additional retrieval operations can increase processing time.

The design separates a persistent historical structure from a temporary working representation, which is the separation on which the present account depends: the history is the object over which $\sim_{\mathcal{R}}$ is defined, and the workspace is a projection constructed per task. The reported evaluations establish measured performance on specified tasks, not exact sufficiency in the sense of Definition 1. They therefore leave open both whether the working representation preserves every distinction those tasks require and whether it would preserve distinctions made decisive by later revisions. Proposition 2 shows why the second question is real: sufficiency for a task collection need not extend to a larger revision class. The natural test, and the one that distinguishes a thread-based organization from a similarity-based one, is to introduce such a revision and compare whether the newly decisive evidence is recovered. That test measures performance relevant to sufficiency under a member of $\mathcal{R} \setminus T$ and so probes a property the original evaluations cannot, though a finite empirical test cannot establish the factorization property over the whole domain of histories.

8.3 Layered Monitoring and Compressed Context

OpenAI describes a monitoring architecture in which inexpensive activation-based probes screen first, model-based review of behavior follows, and selected cases escalate to investigators able to consult broader execution histories and telemetry [3]. A design decision of particular interest is that an initial reviewer may receive an agent’s own summary of its earlier activity, while later stages may inspect the full context where it is available.

Here the projection is the agent’s summary, and the stages differ in which projection they receive. A first-stage monitor can be entirely correct relative to what it observes and still be unable to determine whether a violation occurred, since the discriminating distinction may lie in the discarded part. Improving the monitor does not restore what the projection removed; only changing what the monitor is given does. The report also notes that conditioning training on monitoring outcomes may teach concealment instead of removing the underlying conduct. In the present vocabulary, distinctions erased by a projection cannot directly affect an objective

computed from it, so conduct that differs from benign conduct only in the erased respect is not penalized by that objective. Whether a system exploits this depends on its optimization dynamics and on the actions available to it. The architectural response, which is to keep a path to the unprojected history and to avoid optimizing directly against the screening layer, can be read, in the vocabulary of this paper, as an attempt to preserve sufficiency of the escalation path while limiting the pressure on the screening layer to degrade it. Whether the attempt succeeds is an empirical question about the system. The effectiveness of the arrangement is a claim of the reporting organization about its own system, and nothing in the present framework bears on whether that claim is correct.

8.4 Tests as a Sample of the Specification

Deng and colleagues construct InfraDev, a benchmark of 36 tasks drawn from real changes to distributed training, inference, and reinforcement-learning infrastructure, with additional hidden tests that enforce requirements the original tests leave unchecked [4]. The hidden tests are reported to reject 32 percent of submissions that pass every official test, and the best of six frontier models is reported to resolve 38.9 percent of tasks.

The correspondence with Proposition 2 requires a construction that should be stated and not assumed. Take $\mathcal{H}$ to be the set of candidate implementations, and treat each test as a Boolean-valued policy on implementations. The official tests form a collection T and the full specification a larger family $\mathcal{R}$. Acceptance by T is a property of a single implementation, namely passing every test in T , whereas $\sim_T$ relates pairs of implementations that no test in T separates; two implementations that fail different tests are rejected alike and may still lie in different classes. Because each test is Boolean-valued, $\sim_T$ is an equivalence. The reported failures show that passing the original tests does not imply passing the enlarged collection. Strict refinement requires a witness, a pair of implementations x, y with $t(x) = t(y)$ for every $t \in T$ and $r(x) \neq r(y)$ for some $r \in \mathcal{R}$, and the existence of such a pair is a premise to be established and not assumed. The domain typically supplies one: an implementation passing everything and another passing the original tests while failing an added test. Given a witness, $\sim_{\mathcal{R}} \subsetneq \sim_T$. Informally, $\text{PassesTests}(x) \not\Rightarrow \text{Admissible}(x)$. The consequence is sharper when the artifact under modification is the infrastructure that evaluates later artifacts, because an error left unchecked by the sample can propagate into later evaluations, depending on which execution paths use the affected infrastructure. The case concerns operational adequacy, since the question is whether the submitted code in fact satisfies requirements that the stated checks did not examine.

9 What Follows

Three consequences are worth stating plainly. First, sufficiency for a tested policy collection does not establish sufficiency for a larger revision class unless the collection is known to induce the same verdict-equivalence relation. The obstacle is not finiteness alone but the absence of a demonstrated separating basis for the admissible revisions, and Proposition 3 states exactly what would have to be shown. Second, the retention an authority must provide depends on the diversity of historical distinctions that its authorized revisions can make, and not on the number or apparent complexity of those revisions, as developed under retention in Section 10. Third, the three adequacies of Section 7 must be examined separately, for the reason given

there.

There is a constructive reading. If the revision class can be characterized, even coarsely, then a conservative sufficient projection can be built from a relation that refines $\sim_{\mathcal{R}}$, generated from the features that class can reference, accepting that it may retain distinctions no admissible revision uses. A relation coarser than $\sim_{\mathcal{R}}$ would instead merge histories that some revision distinguishes, and would not be sufficient.

This suggests a revision-sensitive certificate for a projection: a tuple consisting of the projection π , a stated verdict set $\mathbb{V}_{\pi}$, a stated policy family $\mathcal{R}_{\pi}$ over $\mathbb{V}_{\pi}$, and an independently checkable argument that $\ker \pi \subseteq \sim_{\mathcal{R}_{\pi}}$. A later policy ρ can be admitted to the family the projection supports only after it is shown to factor through π , which by Proposition 1 is the condition $\ker \pi \subseteq \ker \rho$. A policy that fails the condition is not evaluated from the projection; the request is routed to a more informative representation, which may be the authoritative history itself. The certificate gives the central characterization the form of an interface between a retained history and the machinery that evaluates policies against it, and it requires no particular memory architecture. The certificate specifies an obligation and not a universal decision procedure: independent checking is available only where the projection and policy languages admit an effective proof system, or where an externally supplied proof of factorization can be verified. The admission test is developed in Section 6, and admission preserves the certificate: by Proposition 4, if π is sufficient for $\mathcal{R}_{\pi}$ and $\ker \pi \subseteq \ker \rho$, then π is sufficient for $\mathcal{R}_{\pi} \cup \{\rho\}$. The old argument still certifies only the old family. The certificate for the enlarged family is the old argument together with the newly verified inclusion $\ker \pi \subseteq \ker \rho$, since $\ker \pi \subseteq \sim_{\mathcal{R}_{\pi}}$ and $\ker \pi \subseteq \ker \rho$ together give $\ker \pi \subseteq \sim_{\mathcal{R}_{\pi}} \cap \ker \rho = \sim_{\mathcal{R}_{\pi} \cup \{\rho\}}$, and a certificate should retain the evidence for each admission and not merely rename the supported family.

The practical research question is therefore the characterization of revision classes: which classes of policy change admit a finite feature basis, which require the full history, and which can be handled by retaining an unprojected path that is consulted only when a revision falls outside the basis. The last option is the one that layered systems already approximate, and it suggests that a principled treatment would specify, for each projection in use, the revision class for which it is certified and the condition under which a request must be routed to the unprojected record.

10 Practical Consequences

The first consequence concerns retention. Because the cost of a sufficient reduction depends on the range of revisions the authority may adopt and not only on the present policy (Corollary 1), a decision to prune, summarize, or downsample a record should begin from an enumeration of the revision class and not from the current policy. In practice this suggests keeping, for each projection in use, a short manifest stating the features the revision class can reference, the feature basis from which the projection was built, and the rule that routes any request falling outside that basis to the unprojected record. The manifest is inexpensive to maintain and is the operational content of the corollary: it records what the projection is certified against, and therefore what it is not.

The second concerns evaluation. A pass rate is silent on the relation the test collection induces, so a benchmark result is more informative when it states which distinctions the collection separates and for which revision class it is meant to certify. Hidden tests reduce the gap

between $\sim_T$ and $\sim_{\mathcal{R}}$ without closing it unless the enlarged collection is known to be separating, and the defensible claim after adding them is sufficiency for the enlarged collection and not for the specification. By Proposition 3 the operative question is one of coverage of distinctions and not of the count of tests, so a benchmark author should be asked to give an argument that the collection separates the class, in addition to reporting its size. The same reasoning applies to memory systems. A thread-structured memory is evaluated on specified tasks, and the further measurement that probes $\mathcal{R} \setminus T$ is a revision-injection suite, reported separately from the task suite, in which a revision makes previously peripheral material decisive and the question is whether the newly decisive evidence is recovered.

The third concerns monitoring and formalization, where the same point recurs in different forms. A monitor that receives a summary may be correct relative to its input and unable to decide, so the useful design requirement is a path to the unprojected record that is exercised and not merely available, since a path never taken provides no assurance. Where a screening objective is computed from a projection, it is insensitive to the distinctions the projection erased, which is a reason to be cautious about optimizing directly against it. In formalization, a proof checker and an alignment check answer different questions, and a pipeline that treats them as one gate will pass statements that type-check while merging claims the source distinguishes. The alignment check is better run as its own gate and at the level of the project, because constraints that are missing from a statement often sit in declared scope elsewhere.

The fourth concerns institutional design. The retention an authority must provide depends on the diversity of historical distinctions its authorized revisions can make, and not on the number or apparent complexity of those revisions: a very large policy family whose members all factor through a small feature set requires little, while a family that separates many histories requires much. Where the family does separate many histories, the goals of freely revisable policy and cheap records cannot both be had, and the trade should be made explicitly and not discovered when a revision arrives. Constraining the revision class to a fixed vocabulary is a legitimate response, provided the refusal of any revision the vocabulary cannot express is visible in the institution’s accounting. Because the verdict set is part of the specification, an institution that expects to add verdict categories should expect that a policy using them may separate histories its present projections merge.

Taken together these suggest a procedure that can be stated in a sentence: treat the revision class, and not the current policy, as the object that determines what may be discarded, and record for each projection the class it is certified against, the verdicts it is certified for, the separating basis for that certification, and the condition under which the unprojected record must be consulted. The three adequacies are then examined separately, so that a strong result on one is not taken for a result on another.

11 Limits of the Account

The characterization is deliberately abstract. It treats verdicts as functions of whole histories and ignores the computational cost of evaluating $\sim_{\mathcal{R}}$; for a rich revision class, that cost may be no easier than the original decision problem. It treats the revision class as given, whereas in practice the difficulty lies largely in saying what the authority could do. It also assumes $\mathcal{R}$ is fixed in advance; a class that can expand, as when an authority gains new powers, is not covered, and the cost bound of Corollary 1 then understates the true cost.

It does not address the case in which the revision changes the notion of history itself,

for example by requiring observation of a quantity that was never recorded; in that case no projection of the existing record can be sufficient, and the matter is one of data collection, not of compression. Formally, if histories are produced from world states by a recording map $o : W \rightarrow \mathcal{H}$ from world states W , a world-level policy $q : W \rightarrow \mathbb{V}$ is evaluable from $\pi \circ o$ exactly when $\ker(\pi \circ o) \subseteq \ker q$, so a log that never recorded the provenance of instructions cannot support a provenance-sensitive policy however well it is preserved. An append-only history prevents the later destruction of recorded distinctions; it cannot supply distinctions that were never observed. There are thus two sources of irrecoverable loss, the recording process and the subsequent projection. Because $\ker o \subseteq \ker(\pi \circ o)$, a distinction lost in recording remains lost under every projection: if $\ker o \not\subseteq \ker q$ then $\ker(\pi \circ o) \not\subseteq \ker q$, and no improvement in retention or projection design can make such a policy evaluable. And the reading of the four recent works is a reading of reported results through a lens they were not built to fit; where those reports make empirical claims, the claims belong to their authors.

References

- [1] H. Shen, R. Qiu, L. Zhi, and X.-S. Gao. MechAlign: Evaluating Semantic Alignment at the Project Level in Mathematical Formalization. arXiv:2610.35269, October 2026.
- [2] M. Zhang et al. TrawMem: Reframing Agent Memory as Task-Oriented Role-Adaptive Workspaces. arXiv:2610.36918, October 2026.
- [3] OpenAI. How We Detect Frontier Model Misalignment. Technical report, October 29, 2026.
- [4] Y. Deng et al. InfraDev: Can Coding Agents Develop AI Infrastructure, and How Can We Tell? arXiv:2610.36589, October 2026.
- [5] H. Yang, J. Chen, and Y. Gao. OA-RAG: Ontology-Based Agentic RAG for Faithful LLM Reasoning. arXiv:2610.36335, October 2026.
- [6] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.
- [7] P. R. Halmos and L. J. Savage. Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. *Annals of Mathematical Statistics*, 20(2):225–241, 1949.
- [8] E. L. Lehmann and H. Scheffé. Completeness, Similar Regions, and Unbiased Estimation: Part I. *Sankhyā*, 10:305–340, 1950.
- [9] D. Blackwell. Equivalent Comparisons of Experiments. *Annals of Mathematical Statistics*, 24(2):265–272, 1953.
- [10] A. Nerode. Linear Automaton Transformations. *Proceedings of the American Mathematical Society*, 9(4):541–544, 1958.
- [11] A. Nash, L. Segoufin, and V. Vianu. Views and Queries: Determinacy and Rewriting. *ACM Transactions on Database Systems*, 35(3):21, 2010.

- [12] N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.
- [13] E. Clarke, O. Grumberg, S. Jha, Y. Lu, and H. Veith. Counterexample-Guided Abstraction Refinement. In *Computer Aided Verification (CAV 2000)*, Lecture Notes in Computer Science 1855, pages 154–169, 2000.