CITIUM
Verification Preprint

Ian Broderick
Citium · editor@citium.ovrinlabs.ca
ORCID 0009-0009-7198-0679

Structural Inevitability: Specification-Side Source Disambiguation

for Oracle-Limited Verification

ABSTRACT

Compression rewards explanations that make efficient use of the evidence it receives. The difficulty studied here appears when the verifier's evidence is too narrow: a false completion can satisfy the accessible checks and still cost no more, structurally, than the true one.

This paper isolates that regime. A budgeted post-hoc verifier observes an output through a feasible check family and the induced observation vector. If a true output and a coherent false output induce the same observation vector, every verifier restricted to that channel must treat them identically; simultaneous soundness and completeness fail on any candidate class containing both. This is a limitation of a fixed information channel, not a general impossibility result for verification. Here, oracle-limited verification means verification constrained by the evidence channel available after generation.

Problem-Solution Isomorphism / Structural Inevitability (PSI/SI) moves part of the control surface upstream. Before inference, it adds source hierarchy, scope, invariants, admissible and invalid methods, success criteria, and output contracts. The aim is not longer prompting; it is a margin: the intended semantic class should become uniquely minimal under an ideal or proxy structural score by enough distance to survive approximation error. The scale claim is therefore conditional. Recovery improves with capability only when that margin is preserved and approximation error decreases.

Keywords: structural inevitability, PSI/SI, Kolmogorov complexity, Solomonoff induction, verification, specification completeness, AI safety

INTRODUCTION

Compression has no private access to truth. It works from a source, a representation, and the evidence supplied by the task. When those elements make the true structure simpler than its rivals, compression and truth align. When they leave a coherent false structure equally cheap to describe, the candidates are structurally tied. The second regime is the subject of this paper.

Solomonoff induction formalizes simplicity-biased prediction (Solomonoff, 1964a; Solomonoff, 1964b; Hutter, 2005), and MDL formalizes related description-length principles for model selection (Rissanen, 1978; Grunwald, 2007). Modern language modeling is closely connected to compression through predictive coding length (Deletang et al., 2024). In many domains, the true explanation is the simplest and most predictive one. The narrower point for verification is conditional: truth is favored when the task specification and evidence make the true completion structurally cheaper than coherent false alternatives.

Recent work motivates this formulation. Krestnikov (2026) argues that next-token prediction rewards internally consistent hypotheses when they compress the data efficiently, even when those hypotheses are false. Young and Witbrock (2024) frame transformers as approximations of Solomonoff induction, and Wan and Mei (2025) study large language models as computable approximations to Solomonoff-style prediction. These papers support the modeling assumption; they do not establish that deployed models uniformly approximate Solomonoff predictors on arbitrary task-specific candidate classes.

Post-hoc verification remains essential when it supplies an independent oracle: proof checking, hidden tests, executable specifications, physical measurement, high-quality retrieval, or expert review can decisively separate true and false outputs. The hard case arises when the verifier has access only to checks that the coherent false completion can also satisfy. In that case, the difficulty is located in the information channel available after generation.

PSI/SI places the disambiguating work before inference. Instead of asking a weak downstream channel to recover a distinction the source never made, it strengthens the specification with authority, scope, invariants, method boundaries, and success conditions. Structural inevitability is reached only when the intended semantic class has a margin over its coherent rivals large enough to survive proxy error and representation tolerance.

The specification condition itself is scale-independent. The favorable scale claim is narrower: if the specification creates a positive structural margin and approximation error decreases with capability, the recovery bound for the intended completion improves.

What this paper does not claim. The result does not say post-hoc verification is broadly weak. Exact proof checkers, hidden tests, executable specifications, measurements, and independent experts can be decisive when they expose distinctions coherent false completions cannot satisfy. The paper also does not claim that exact Kolmogorov complexity is computable, that real model capability always reduces approximation error, or that source disambiguation automatically creates a margin in every domain.

Contributions. The contribution is fourfold.

1. It replaces the blanket claim that post-hoc verification is scale-negative with a conditional theorem about observation-channel indistinguishability.
2. It separates practical difficulty, computational difficulty, epistemic indistinguishability, and structural impossibility under equal-complexity hypotheses.
3. It states the PSI/SI guarantee as margin-stable recovery under proxy error, explicitly marking fixed representation, semantic equivalence classes, positive margin, and approximation assumptions.
4. It turns the inversion into a falsifiable empirical program.

RELATED WORK AND SCOPE

Algorithmic Probability, MDL, and Compression

The formal ancestry is already established. Solomonoff induction, AIXI, Kolmogorov complexity, and MDL formalize simplicity-biased prediction (Solomonoff, 1964a; Solomonoff, 1964b; Hutter, 2005; Li and Vitanyi, 2008; Rissanen, 1978; Grunwald, 2007). The new step is a design consequence for AI verification: if the intended completion can be made conditionally simpler than coherent false alternatives by a margin that survives approximation error, increasing capability can improve reliability even as generated outputs become more sophisticated.

Post-Hoc Verification and Alignment

RLHF, Constitutional AI, debate, scalable oversight, and model critique all shape behavior after or around generation (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Irving et al., 2018; Bowman et al., 2022). PSI/SI is not a substitute for that stack. Its claim is upstream: output-level procedures receive whatever ambiguity the source and task channel leave unresolved, while specification work tries to change the candidate landscape before generation begins.

Formal Specification and Verification

Formal methods already demonstrate the power of specifying requirements, invariants, and interfaces before verification (Meyer, 1988; Lamport, 2002; Jackson, 2011; Katz et al., 2017; Liu et al., 2021). PSI/SI complements formal verification with a graded, information-theoretic account of why specification strengthening helps even in partially formalized settings. Formal verification checks that an artifact satisfies a property; PSI/SI asks whether the specification is complete enough that the intended artifact is the uniquely simplest completion. The observation-channel lemma uses an elementary transcript argument: once two cases induce the same observable transcript, no decision rule over that transcript separates them.

THE POST-HOC VERIFICATION PROBLEM

The formalization below separates the output from the verifier's access to the output. A verifier may apply every check feasible under its budget. Those checks produce an observation vector. The decisive case is a pair of outputs with different truth values and the same vector. Once that happens, the verifier's decision rule receives identical input for both cases.

This is the narrow sense in which post-hoc verification becomes structurally limited: the limitation belongs to the information made available to the verifier after generation.

Let x be a task specification, w a latent world state or intended semantics, and $R(w, x, y) \in \{0, 1\}$ the correctness relation for an output y. A post-hoc verifier leaves x unchanged before generation. It observes an output y and applies a feasible set of checks after the fact.

DEFINITION 3.1 (BUDGETED POST-HOC VERIFIER)

Fix a verification budget B. Let T_B be the set of checks feasible under that budget: human review, secondary model judgment, retrieval, tool calls, test suites, proof checkers, measurements, or other procedures. The verifier observes

$$o_B(x, y) = (t(x, y))_{t \in T_B}.$$

A deterministic budgeted post-hoc verifier is any decision rule

$$V_B(o_B(x, y)) \in \{\text{accept}, \text{reject}, \text{abstain}\}.$$

A randomized verifier is a distribution over these verdicts conditioned on $o_B(x, y)$. The verifier is post-hoc when the source specification x is not enriched by the verifier before generation.

DEFINITION 3.2 (COHERENT FALSE COMPLETION)

A false completion $\tilde{y}$ is coherent relative to (x, w, B, L, η) when $R(w, x, \tilde{y}) = 0$, $\tilde{y}$ satisfies the local signals available to T_B, and there exists a true completion $y^\star$ such that

$$L([\tilde{y}]_x \mid x) \le L([y^\star]_x \mid x) + \eta,$$

where L is a declared proxy structural score and $[\cdot]_x$ denotes the task-relevant semantic class. The false output is therefore wrong, locally plausible, and structurally competitive under the information available to the generator or verifier.

Worked Examples

Rule-world verification. Consider a synthetic rule task where the public source says that items a and b are accepted and item c is rejected. The hidden intended rule is: accept an item iff it has tag T and is not exempt. A coherent false rule says: accept an item iff it has tag T. If the public examples contain no exempt item, a verifier restricted to public checks observes

$$o_B(x, y^\star) = (1, 1, 1) = o_B(x, \tilde{y}).$$

The two rules differ on a private discriminator d that has tag T and is exempt: $y^\star(d) = 0$ while $\tilde{y}(d) = 1$. No verifier whose decision is a function only of the public vector can accept the true rule and reject the false rule. Source enrichment breaks the tie by adding the missing invariant: exempt overrides tag membership. Under a rule-grammar proxy, the operational margin is the gap between the enriched true rule and the cheapest admissible rival after that invariant is present.

Public-test code generation. Consider a code task asking for a shipping-cost function. The public tests cover domestic packages below two kilograms and international packages below two kilograms. The intended semantics also say that packages above two kilograms require freight routing. A coherent false implementation passes every public test by ignoring the freight boundary. A post-hoc verifier that runs only public tests sees the same pass vector for the true implementation and the false implementation. A stronger verifier with hidden overweight cases separates them; PSI/SI attempts to move that separation upstream by making the freight boundary, units, invalid inputs, and abstention behavior part of the source before generation.

Four Bottlenecks

Post-hoc verification can fail for distinct reasons. These failure modes differ in repair strategy, so they should remain separated.

Practical difficulty.
The answer may be checkable in principle while exceeding realistic budgets of time, expertise, context, or attention.

Computational difficulty.
The object may be formalized, but checking it or obtaining a checkable certificate may be expensive.

Epistemic indistinguishability.
The accessible evidence leaves correctness unsettled.

Structural equal-complexity.
The true and false hypotheses are both compatible with the evidence, and the false one is no more expensive to encode. A compression-biased generator lacks a structural reason to prefer the true hypothesis, and a verifier restricted to the same information lacks the evidence needed to recover truth by inspection alone.

LEMMA 3.3 (OBSERVATION-CHANNEL INDISTINGUISHABILITY)

Fix a task x, latent semantics w, and verification budget B. If there exist outputs $y^\star$ and $\tilde{y}$ such that

$$R(w, x, y^\star) = 1, \qquad R(w, x, \tilde{y}) = 0, \qquad o_B(x, y^\star) = o_B(x, \tilde{y}),$$

then every deterministic verifier restricted to o_B gives both outputs the same verdict. Every randomized verifier assigns them the same acceptance probability. Here soundness means no false candidate in the class is accepted, and completeness means the true candidate in the class is accepted. Simultaneous soundness and completeness are impossible on any candidate class containing both outputs.

PROOF.

Any budgeted verifier is a function, or distribution-valued function, of $o_B(x, y)$. If the true and false outputs induce the same observation vector, the verifier receives identical input in both cases. Equal input gives equal verdicts or equal verdict distributions. Accepting both is unsound; rejecting both is incomplete; abstaining on both is nondecisive on a class containing both.

DEFINITION 3.4 (ORACLE-LIMITED TASK FAMILY)

A task family F is oracle-limited relative to budget B, verifier policy class Π_B, structural proxy L, and tolerance η when there is a subfamily $F_0 \subseteq F$ that is either designated as the regime of interest or has positive measure $\mu(F_0) > 0$ under a fixed task distribution μ, such that for each $x \in F_0$ there exist latent semantics w, a true output $y^\star$, and a false output $\tilde{y}$ with

$$R(w, x, y^\star) = 1, \qquad R(w, x, \tilde{y}) = 0,$$

$$L([\tilde{y}]_x \mid x) \le L([y^\star]_x \mid x) + \eta,$$

and for the deployed verifier policy class, the feasible post-hoc observations fail to expose the relevant distinction. In the strongest form, $o_B^\pi(x, y^\star) = o_B^\pi(x, \tilde{y})$ for every $\pi \in \Pi_B$, where o_B^π denotes the observation transcript exposed under verifier policy π. The limitation is relative to B, Π_B, and the observation channel; a larger budget or independent oracle may remove it.

DEFINITION 3.5 (RESIDUAL FALSE-PASS SET)

Let $G_c(x)$ be the set of outputs reachable by a capability-c generator. For capability level c, define the residual false-pass set

$$FP_B(c; x) = \{\tilde{y} \in G_c(x) : R(w, x, \tilde{y}) = 0,\ V_B(o_B(x, \tilde{y})) = \text{accept}\}.$$

PROPOSITION 3.6 (MONOTONE WORST-CASE FALSE-ACCEPTANCE RISK UNDER FIXED CHANNEL)

Fix x, w, B, o_B and a verifier decision rule V_B. Suppose $G_c(x) \subseteq G_{c'}(x)$ for $c < c'$. Define the worst-case false-acceptance risk

$$\mathrm{FAR}_B^{\max}(c; x) = \sup_{\tilde{y} \in G_c(x):\, R(w,x,\tilde{y})=0} \Pr[V_B(o_B(x, \tilde{y})) = \text{accept}],$$

with $\sup \varnothing = 0$. Then $\mathrm{FAR}_B^{\max}(c; x) \leq \mathrm{FAR}_B^{\max}(c'; x)$. Distributional false acceptance may still improve if generator probability mass moves away from the residual false-pass set.

PROOF SKETCH.

The supremum is taken over reachable false candidates, and that set is preserved or enlarged as capability increases while the verifier channel and decision rule remain fixed. A supremum over a superset cannot be smaller. Lemma 3.3 explains why channel-indistinguishable false outputs cannot be removed by a decision rule over the same observation vector.

REMARK 3.7 (SCOPE OF THE PROPOSITION)

The proposition applies to bounded, oracle-limited, one-shot or weakly iterative post-hoc verification against coherent false completions. It leaves room for verifiers that improve in absolute accuracy, false outputs that remain detectable, and verification systems with exact or independent checkers. In those cases, the bound can be broken by increasing the information channel.

STRUCTURAL INEVITABILITY

The previous section treats ambiguity as a property of the verifier's channel. Structural inevitability treats ambiguity as a property of the source presented to the generator. The source can be strengthened before inference by specifying authority, scope, invariants, admissible methods, excluded methods, output contracts, and success criteria. The target condition is a positive conditional description-length gap between the intended semantic class and every coherent rival.

DEFINITION 4.1 (SEMANTIC SOLUTION CLASSES)

For a problem P, let $Y(P)$ denote the surface response space and let $\equiv_P$ be a task-relevant equivalence relation supplied by the specification or charter. Let $S(P) \subseteq Y(P)/\equiv_P$ denote admissible semantic solution classes. Raw string variants, irrelevant paraphrases, formatting changes, and other surface differences are quotiented away.

DEFINITION 4.2 (STRUCTURAL INEVITABILITY)

A problem specification P achieves structural inevitability for $S^* \in S(P)$ with margin $\delta > 0$ when

$$K_U(S^* \mid P) + \delta \leq K_U(S \mid P) \qquad \text{for every } S \neq S^* \in S(P),$$

relative to a fixed reference representation U or code family.

REMARK 4.3 (REPRESENTATION AND COMPUTABILITY)

Exact Kolmogorov complexity is incomputable and machine-invariant up to an additive constant. PSI/SI uses K_U as an ideal analytical object. Operational claims must be stated relative to fixed computable proxies--for example proof length, program length, negative log-likelihood under a frozen reference model, verifier certificate length, grammar-MDL, or domain-specific description length. If $\hat{\delta}_L(P)$ is the estimated proxy margin and $\hat{\epsilon}_L(P)$ is the proxy error estimate, an operational SI claim requires a margin threshold such as

$$\hat{\delta}_L(P) > 2\hat{\epsilon}_L(P) + \tau_U,$$

where τ_U represents representation tolerance.

DEFINITION 4.4 (PSI/SI PROTOCOL)

The PSI/SI protocol searches the remaining ambiguity in P and adds only fields that reduce it: authority order, definitions, invariants, permitted and excluded methods, output contracts, and verification criteria. A domain instance is treated as provisionally complete when the target class is estimated to retain a positive margin against coherent alternatives.

Operational use is diagnostic. If a rival class remains cheap, the question becomes which absent constraint is preserving the tie: authority, temporal scope, invariant, method boundary, schema, or abstention rule. A field earns its place only if it changes the relative structural cost of the admissible classes.

THEOREM 4.5 (MARGIN-STABLE SI RECOVERY UNDER PROXY ERROR AND OPTIMIZER TOLERANCE)

Let $S(P)$ be finite and let $L^*(S \mid P)$ be an ideal structural score over $S(P)$. Suppose S^* has margin $\delta > 0$:

$$L^*(S^* \mid P) + \delta \leq L^*(S \mid P) \qquad \forall S \neq S^*.$$

Let L_θ be a computable proxy satisfying

$$\sup_{S \in S(P)} |L_\theta(S \mid P) - L^*(S \mid P)| \leq \varepsilon_\theta.$$

Let $\hat{S}_\theta$ be an η-optimal proxy minimizer:

$$L_\theta(\hat{S}_\theta \mid P) \leq \min_{S \in S(P)} L_\theta(S \mid P) + \eta.$$

Here $\tau \geq 0$ is an additive representation tolerance in the structural score. Call a rival S τ-distinguishable from S^* when $L^*(S \mid P) \geq L^*(S^* \mid P) + \tau$; the recovery claim excludes only τ-distinguishable rivals. If $\delta > 2\varepsilon_\theta + \eta + \tau$, then no η-optimal proxy minimizer can select a rival outside the declared τ-tolerance; equivalently, recovery is unique modulo that tolerance. With $\eta = \tau = 0$, exact proxy minimization recovers S^* whenever $2\varepsilon_\theta < \delta$.

PROOF SKETCH.

For every rival S, the proxy score of S^* is at most $L^*(S^* \mid P) + \varepsilon_\theta$, while the proxy score of S is at least $L^*(S \mid P) - \varepsilon_\theta$. The ideal margin gives $L^*(S \mid P) \geq L^*(S^* \mid P) + \delta$. Thus $L_\theta(S \mid P) - L_\theta(S^* \mid P) \geq \delta - 2\varepsilon_\theta$. If $\delta > 2\varepsilon_\theta + \eta + \tau$, every rival outside the declared τ-tolerance has proxy score more than η above S^*, so no η-optimal proxy minimizer can select it.

COROLLARY 4.6 (CONDITIONAL MONOTONE RECOVERY UNDER DECREASING PROXY ERROR)

Under the proxy-error setup of Theorem 4.5, let c be a capability index. Suppose $S(P)$, δ, β, and τ are fixed, proxy error $\varepsilon(c)$ is nonincreasing, and optimizer tolerance $\eta(c)$ is nonincreasing. Define

$$m_c = \delta - 2\varepsilon(c) - \eta(c) - \tau.$$

If, for the same capability-indexed proxy $L^{(c)}$, a model samples semantic classes with Gibbs weight proportional to $\exp(-\beta L^{(c)}(S \mid P))$, then whenever $m_c > 0$,

$$\Pr_c(S^* \mid P) \geq \frac{1}{1 + (|S(P)| - 1)\exp(-\beta m_c)}.$$

If m_c is nondecreasing with capability, this lower bound is nondecreasing. This is the precise conditional sense in which source-side structural margin can improve with capability.

THE INVERSION

The inversion is about where information enters the process. A stronger generator searches a larger and more coherent region of the output space. If the verifier's channel is fixed, that search can produce false completions that satisfy all visible checks. If the source has already been disambiguated, the same increase in capability works in the other direction: the target is shorter to describe relative to the enriched specification, and a better approximation to the conditional predictor concentrates on it.

The crossover in Figure 1 depends on two measurable quantities: the residual false-pass set for the post-hoc verifier and the estimated structural margin induced by the specification. PSI/SI can outperform post-hoc checking in regimes where the margin is positive and approximation error remains below the margin-relevant threshold. Exact or cross-family oracles can move the verification curve upward; incomplete or unstable charters can move the PSI/SI curve downward.

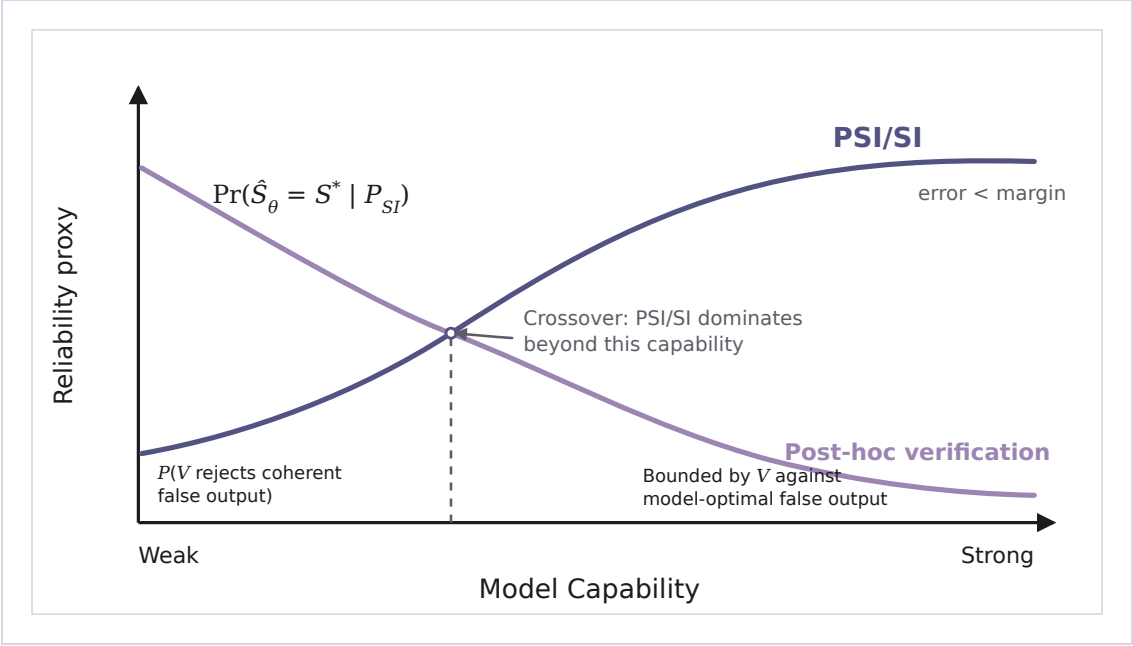

Figure 1. Schematic conditional regime diagram, not empirical data. Here P_{SI} denotes the PSI/SI-enriched source. In the oracle-limited regime, post-hoc verification can become weaker relative to coherent false outputs as model capability increases while the verifier channel remains fixed. PSI/SI becomes stronger only when source disambiguation gives the intended completion a positive operational margin.

EMPIRICAL VALIDATION PROGRAM

The theory should be tested as an interaction hypothesis. The main empirical question is: as generator capability increases, do realistic post-hoc verifiers fail to improve fast enough on coherent false outputs, and does structured pre-inference disambiguation improve more with capability than looser forms of context augmentation?

Operational Variables

Model capability.
A task-external capability index, such as an item-response or standardized aggregate score on a held-out battery, with parameter count and inference compute treated as secondary proxies.

Verification effectiveness.
AUROC, balanced accuracy, false acceptance rate, and false rejection rate, reported both on all outputs and on the coherent-false subset.

Specification completeness.
The weighted fraction of ambiguity-bearing fields disclosed before generation: source hierarchy, temporal scope, admissible operations, exception ordering, invariants, schema, and abstention rules.

PSI/SI conditioning.
A structured pre-inference specification that marks authority, scope, invariants, success criteria, and output contracts before the model answers.

Experimental Arms

A useful study uses three generation arms and multiple verifier regimes:

1. minimal prompt with minimal disambiguation;
2. retrieval-augmented or context-augmented prompt with more evidence but weak structure;
3. charter/specification-conditioned prompt with explicit invariants, source hierarchy, and success criteria.

Verifier regimes should include no verifier, a weak verifier, a strong cross-family verifier, and an exact oracle where available. Domains should include mathematics, theorem proving, code generation with hidden tests, scientific question answering with known ground truth, adversarial synthetic worlds, and rule-based legal or policy reasoning.

Minimum Publishable Experiment

A first experiment can be synthetic and fully auditable. Generate rule-world tasks with a hidden source rule, a coherent false rule that matches all public examples, and a private discriminator check. Compare three generation arms: minimal prompt, token-matched unstructured context, and PSI/SI-conditioned specification with source authority, rule scope, exception order, output contract, and abstention rule. Compare three verifier regimes: no verifier, public-check verifier, and private oracle verifier. Report correctness, false acceptance on coherent false outputs, invariant satisfaction, abstention accuracy, and proxy margin $\hat{\delta}_L$. The hypothesis is falsified if PSI/SI conditioning fails to improve over token-matched context, or if proxy margin fails to correlate with correctness.

Falsification Conditions

The inversion is weakened if exact or cross-family post-hoc verification scales well enough to erase coherent-false false acceptance, or if structured PSI/SI conditioning fails to improve more with capability than ordinary retrieval/context augmentation. It is strengthened if PSI/SI conditioning increases invariant satisfaction and correctness while the residual coherent-false pass rate remains flat or worsens for budget-limited post-hoc verification.

DISCUSSION

PSI/SI is a specification-first design principle with an ideal information-theoretic limit. The practical question is simple to state: how much ambiguity can be removed before inference, and which domains admit measurable margins? RLHF, Constitutional AI, process supervision, retrieval, interpretability, formal verification, and guardrails remain part of the engineering stack. They supply constraints, evidence, verifiers, and escalation paths. In the present framing, they improve either the source before generation or the independent channel after generation.

Limitations

1. **Exact K computation.** Exact Kolmogorov complexity is incomputable. Operational work requires declared proxies.
2. **Representation dependence.** Margins must dominate encoding constants or be stated relative to a fixed representation.
3. **Independent oracles.** Exact or independent oracles can make post-hoc verification strong.
4. **Specification algorithm boundary.** The inevitability frontier identifies useful constraints while falling short of a universal search procedure.
5. **Domain knowledge.** Productive specification requires domain expertise and iterative refinement.
6. **Empirical status.** The inversion is experimentally investigable and requires empirical testing beyond the formal apparatus.

CONCLUSION

The claim is narrow and operational. Post-hoc verification is powerful when it has information that separates true completions from coherent false completions. In oracle-limited regimes, a verifier restricted to an observation channel that gives both completions the same vector lacks the information needed to recover the distinction by decision rule alone. PSI/SI changes the channel by enriching the specification before inference.

When that enrichment creates a positive structural margin, capability helps: a better compressor is more likely to recover the inevitable semantic class. The resulting research program is concrete: identify domains with independent post-hoc oracles, identify domains where bounded verification remains oracle-limited, build specifications that create measurable margins, and test whether charter-conditioned generation converts ambiguous tasks into structurally inevitable ones.

REFERENCES

1. Yuntao Bai et al. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073, 2022. doi:10.48550/arXiv.2212.08073.
2. Samuel R. Bowman et al. Measuring progress on scalable oversight for large language models. arXiv:2211.03540, 2022. doi:10.48550/arXiv.2211.03540.
3. Paul F. Christiano et al. Deep reinforcement learning from human preferences. NeurIPS 30, 2017. arXiv:1706.03741. doi:10.48550/arXiv.1706.03741.
4. Gregoire Deletang et al. Language modeling is compression. ICLR, 2024. arXiv:2309.10668. doi:10.48550/arXiv.2309.10668.
5. Peter D. Grunwald. The Minimum Description Length Principle. MIT Press, 2007.
6. Marcus Hutter. Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, 2005.
7. Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. arXiv:1805.00899, 2018. doi:10.48550/arXiv.1805.00899.
8. Daniel Jackson. Software Abstractions: Logic, Language, and Analysis. Revised edition, MIT Press, 2011.
9. Guy Katz et al. Reluplex: An efficient SMT solver for verifying deep neural networks. CAV, pages 97--117, 2017. arXiv:1702.01135. doi:10.1007/978-3-319-63387-9_5.
10. Konstantin Krestnikov. Truth as a Compression Artifact in Language Model Training. arXiv:2603.11749, 2026. doi:10.48550/arXiv.2603.11749.
11. Leslie Lamport. Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers. Addison-Wesley, 2002.
12. Ming Li and Paul Vitanyi. An Introduction to Kolmogorov Complexity and Its Applications. Third edition, Springer, 2008.
13. Changliu Liu et al. Algorithms for verifying deep neural networks. Foundations and Trends in Optimization, 4(3--4):244--404, 2021. doi:10.1561/2400000035.
14. Bertrand Meyer. Object-Oriented Software Construction. Prentice Hall, 1988.
15. Long Ouyang et al. Training language models to follow instructions with human feedback. NeurIPS 35, 2022. arXiv:2203.02155. doi:10.48550/arXiv.2203.02155.
16. Jorma Rissanen. Modeling by shortest data description. Automatica, 14(5):465--471, 1978. doi:10.1016/0005-1098(78)90005-5.
17. Ray J. Solomonoff. A formal theory of inductive inference. Part I. Information and Control, 7(1):1--22, 1964. doi:10.1016/S0019-9958(64)90223-2.
18. Ray J. Solomonoff. A formal theory of inductive inference. Part II. Information and Control, 7(2):224--254, 1964. doi:10.1016/S0019-9958(64)90131-7.
19. Jun Wan and Lingrui Mei. Large language models as computable approximations to Solomonoff induction. arXiv:2505.15784, 2025. doi:10.48550/arXiv.2505.15784.
20. Nathan Young and Michael Witbrock. Transformers as approximations of Solomonoff induction. arXiv:2408.12065, 2024. doi:10.48550/arXiv.2408.12065.