
Ω -Certifiability: A Theory of Certified Ignorance

Eduardo Perez Denadai
Optimality.ai

Abstract

Standard scalar accuracy fails to distinguish between a model that overclaims by answering every prompt and one that underdelivers by systematically declining; both can achieve misleadingly high scores while failing under operational constraints. A learning system is knowledgeable not in proportion to what it answers but in proportion to what it can answer *and* refuse, each with certainty—what we call *certified ignorance*: “I do not know, and here is why,” as a checkable statement rather than a hedge. We introduce Ω -certifiability: the (precision, recall) pair of a policy measured against a pinned formal authority Ω , where perfect (1, 1) means the policy claims exactly the decidable tasks and certifies every claim it makes—no false abstention, no overclaim, measured, not asserted. We prove when a learner can reach that point and when it provably cannot. The reachability conditions form an axiom: conditions C1–C6 (realizability, oracle-derived target, support coverage, visibility, loss alignment, oracle hygiene), supported by two formal postulates (Information and Retention) whose nulls were registered in advance and verified through pre-registered experiments. Empirical evaluations confirm the theoretical predictions: on a frozen benchmark with a pinned oracle, swapping the teaching target from self-imitation to oracle-derived moves yield $38.6 \rightarrow 79.5 \rightarrow 95.5\%$ and *simultaneously* moves overclaim on the undecidable side $58.3 \rightarrow 86.7 \rightarrow 30.0\%$ —both directions registered before the run and replicated at 98.3%; a shuffled-target placebo collapses to 1.7%; five-seed information-preserving interventions are null; a decode-time gate drives overclaim to 0% for every proposer at threshold $\tau \geq 0.9$. We show that the operational limits are structural theoretical bounds rather than engineering artifacts: exact (1, 1) requires the boundary to be a deterministic function of the input (the Chow frontier), off-fragment certifiability is bounded by Rice-type undecidability—abstention outside the fragment is theorem-forced, not prudent—and a measured coverage ceiling of 0 of 2,148 abstentions marks where no rung can help. We close by running the formal protocol for unbiased ground-truth evaluation: on an external propositional-entailment corpus with published labels, the pair measures (1, 1) on the addressable population with zero false certificates—in two logics—under three-layer cross-validation, against the falsifier registered in advance.

1 Introduction

Language models now emit artifacts that carry logical claims—code with assertions, contracts with numbers, actions with preconditions. The bottleneck is no longer generation; it is *trust*: admitting probabilistically produced claims under a mathematically auditable standard. Two failure modes dominate. A model that always answers overclaims; a model that always declines underdelivers. Aggregate accuracy hides both.

This paper is about knowledge, not scorekeeping. Our thesis: *certified ignorance*—the ability of a learner to state, with certainty, exactly what it cannot prove—is the property worth measuring, and it is a property of a *pair*: a learner and a declared fragment of the world it may reason in.

Two classical facts frame the problem and, read correctly, derive it rather than block it. Gödel’s incompleteness says a sufficiently strong, consistent, recursively axiomatized system cannot certify its own consistency from within (?); it does *not* block certainty about particular ignorance—“ φ is not provable in F ” is itself often provable in a stronger system—so self-verification is circular (?) while *external* metaknowledge of one’s own limits remains tractable. Rice’s theorem says every non-trivial semantic property of programs is undecidable (?); so refusing outside a decidable fragment is not prudence but the only sound answer. A theory of certified ignorance must therefore be relative, external, and—for it to be science rather than posture—measured. A first pilot of the loop this paper formalizes shows what measurement buys: declining tasks outside the provable fragment moved from 4% to 58% after training against authority-derived targets, every refusal adjudicated by the authority rather than self-reported. The learner learns the boundary; the authority makes the boundary checkable.

We measure the opposite of accuracy. Fix a deterministic verification oracle Ω —an exact checker whose soundness is policed, not assumed. Let A be the set of tasks whose contracts Ω can certify non-vacuously, and U the rest. A policy π either proposes an artifact or abstains ($\perp$). We say π *claims* on x when $\pi(x) \neq \perp$ and *certifies* when Ω accepts its artifact. The distinction that organizes the paper is drawn here: reported confidence is a property of the model’s *output* and inherits every bias of the reporter, whereas the claim/refuse decision evaluated in this paper is adjudicated by an *external, executable* check—the refusal signal is never produced by the thing being refused about.

Definition 1 (Ω -certifiability). *The Ω -certifiability of π is the pair*

$$\text{prec}(\pi) = \frac{\mathbb{E}[\text{cert}]}{\mathbb{E}[\text{claim}]}, \quad \text{rec}(\pi) = \frac{\mathbb{E}[\text{cert} \cdot \mathbf{1}_A]}{P(A)},$$

where $\text{claim}(\pi, x) = \mathbf{1}[\pi(x) \neq \perp]$ and $\text{cert}(\pi, x) = \mathbf{1}[\Omega(\pi(x)) = \text{CERTIFIED}]$ (zero on $\perp$). If $\mathbb{E}[\text{claim}] = 0$ the first ratio is 0/0: we declare it vacuous ($\text{prec} = \perp$, never 1), so the *abstain-everywhere* policy scores $(\perp, 0)$ and can never be reported as perfect. Perfect Ω -certifiability is $(1, 1)$: $\forall x : \text{claim} = \mathbf{1}_A$ and $\forall x \in A : \text{cert} = 1$.

A policy that abstains everywhere has vacuous precision and zero recall; one that always claims has high recall and collapsing precision. The pair separates them where no scalar can. The metric is *oracle-relative*: without a pinned Ω it is undefined, which is why every reported number in this paper carries an authority fingerprint.

Contributions. (1) A formal definition of Ω -certifiability and its properties (§??); (2) an axiom of reachability—conditions C1–C6 plus the information and retention postulates C7–C8 and the locus principle—with a convergence theorem and the characterized consequence of violating each condition (§??, §??); (3) structural limits as theorems: the imitation fixed point, the Chow frontier ceiling, and the Rice wall of the un-fragmented problem (§??); (4) a controlled experimental program on a frozen benchmark with a pinned oracle: the causal contrast on the teaching target, a pre-registered shuffled placebo, five-seed nulls, machine-teaching sets, a runtime gate, and reward-laundering detection and repair (§??), with the full registration ledger kept as an appendix battery; and (5) the path off oracle circularity, executed: a protocol on an external decidable corpus with published labels, a differential pipeline against third-party artifacts, and a measured $(1, 1)$ with zero false certificates (§??).

2 Related work

Selective prediction. Chow’s reject option (??) predates our setting by five decades: abstain when conditional risk exceeds a threshold. Modern treatments give risk–coverage curves and coverage-optimal selective strategies (????)—including the *consistent selective strategy* of ?, reject unless the whole hypothesis class agrees, which our architecture instantiates: a detached checker re-derives every certificate from the artifact alone and refuses what it cannot discharge. The refusal half of Ω -certifiability is therefore 55-year-old prior art and we claim no novelty for it; our claim is the *training-signal* question the classical literature does not ask—what teaching signal moves the pair, not only what threshold extracts it. Ω -certifiability inherits the reject-option frontier—indeed our ceiling theorem *is* Chow’s rule applied to the posterior of decidability—but differs in three ways: the evaluation target is an executable oracle rather than a label distribution; abstention is *positively* scored when the task is genuinely undecidable; and one asymmetry makes the obligation

Table 1: What each property belongs to. Only the last is measured against a pinned Ω , and only the last is a pair.

property	of whom	meaning
capability	model	correctness on sampled tasks (pass@ k)
verifiability	claim / fragment	whether the claim <i>can</i> be checked at all
calibration	model’s decision	willingness to claim vs. probability of provability
soundness	the oracle Ω	zero false approvals; policed, never assumed
Ω-certifiability	model <i>relative to</i> Ω	the certified pair: precision and recall of claiming

fail-closed rather than risk-tuned: for a certificate the false positive is not an error rate but a transferred liability—someone else will rely on it.

Self-knowledge in language models. Large language models (LLMs) report what they know (?); refusal-aware tuning and uncertainty expression improve self-reports (??). These evaluate against human or judge labels; ours is evaluated by execution of a pinned checker, so both the claim and the refusal are verifiable.

Verifiable feedback and certified learning. Process and outcome verification (??) and reinforcement learning from verifiable rewards (??) use correctness signals to raise pass rates; we score the abstention axis with equal standing and show that a verifier-shaped reward can *launder* failure (§??). Certifiable machine learning (?) trains models with checkable guarantees; Ω -certifiability supplies the operational, oracle-relative metric for claim-vs-decidable behavior and the teaching conditions under which a learner reaches it. The proposer–authority discipline for code claims is developed in companion work (?).

Machine teaching and interference. Machine teaching selects minimal sets that induce a target concept (?); we measure a disagreement-only set beating both a size-matched random set and the full set on the mandate axis. Catastrophic forgetting and gradient-conflict mitigation (???) inform our retention analysis (C8): sequential stages overwrite calibration; joint unions preserve it.

3 Definitions

Setup. Tasks $x \sim D$ over input space $\mathcal{X}$. The oracle Ω is deterministic on artifacts and *sound*: it never certifies a non-conforming artifact. A certificate is *non-vacuous* when it discharges at least one obligation that is not itself an assumed premise (the formal reading of “ Ω certifies x ’s contract”). The decidable set is

$$\mathcal{A} = \{x \in \mathcal{X} \mid \exists \varphi. \Omega(\varphi) = \text{CERTIFIED non-vacuous, } \varphi \text{ encodes } x\text{'s contract}\},$$

$$U = \mathcal{X} \setminus \mathcal{A},$$

and we write $1_{\mathcal{A}}(x) \in \{0, 1\}$; in the main text A and U denote $\mathcal{A}$ and $\mathcal{X} \setminus \mathcal{A}$. A policy $\pi : \mathcal{X} \rightarrow (\text{Artifacts} \cup \{\perp\})$. The oracle-consistent expert π^* emits a certifying φ on A and $\perp$ on U ; it is perfectly Ω -certifiable by construction. The question is never whether a perfect policy exists but whether a *learner* reaches it.

Proposition 1 (Properties). Ω -certifiability (i) is undefined without a pinned oracle; (ii) is monotone non-decreasing in best-of- K sampling on recall with precision held by Ω ; (iii) is improvable by exactly three levers: informative targets (training), a wider oracle (new exact mechanisms), or a different target distribution (which point of the frontier to occupy); and (iv) separates capability (does it solve A ?) from calibration (does it claim only A ?)—the two axes of Definition ??.

The pair is easily confused with its neighbors; Table ?? is the disambiguation. Its practical consequence: a larger-but-noisier model can achieve higher certifiability than a smaller-cleaner one, because the oracle filters the noise while *coverage* is the scarce good; the appropriate comparison axis is certifiability rather than scale.

Effective topology. In the computable topology on $\mathcal{X}$ with Ω a program, $\{x : \text{certifiable}\}$ is Σ_1^0 (effectively open) and $\{x : \text{undecidable}\}$ is Π_1^0 ; both are Borel, so the pair of measures in Definition ?? is well defined. Discovery (Σ_1^0 -witnessing) can only move tasks from U to A ; demotion requires refuting a Π_1^0 statement. The frontier of a benchmark is therefore monotone under discovery—the justification for growing the oracle only by conservative extension.

4 The axiom

The axiom’s spine: *a learner is Ω -certifiable exactly to the extent that the oracle’s boundary is (i) realizable, (ii) taught through an oracle-derived target, and (iii) visible in the input.* Formally:

Assumption 1 (Reachability conditions C1–C6). *C1 **Realizability.** $\pi^* \in \mathcal{H}$ for hypothesis class $\mathcal{H}$.*

*C2 **Target consistency (the crucial one).** The training target is oracle-derived: $T(\cdot | x) = \delta_{\pi^*(x)}$.*

*C3 **Support coverage.** $\text{supp}(D) \subseteq \text{supp}(T)$ and evaluation distribution $= D$. Runtime use: claim iff the posterior estimate $\hat{\pi}_A(x) \geq \tau$ (a supplied boundary).*

*C4 **Visibility.** 1_A is measurable w.r.t. the model’s observed features: $1_A \in \sigma(X)$. We separate (a) information presence $I(1_A; X)$, (b) usability of that information at decode time, and (c) the Rice/halting wall on un-fragmented program text.*

*C5 **Loss alignment.** The empirical risk minimization (ERM) over $\mathcal{H}$ under T is π^* (holds for cross-entropy given C1, C2).*

*C6 **Oracle hygiene.** Ω sound, total on the fragment, deterministic, no FAULT (the halt-on-definite-disagreement alarm); policed by adversarial twins and pinned regressions, never assumed.*

Postulate 1 (Information; C7). *Ω -certifiability improves only in proportion to the information the learning signal carries about 1_A . Information-preserving interventions—paraphrase, reordering, temperature sweeps, resampling—are null by construction; only interventions that add or expose boundary information move the pair.*

Postulate 2 (Retention; C8). *A fixed-capacity policy can exchange behaviours but not add them: sequential training on differently composed distributions overwrites earlier calibration. The effective teaching set is the joint union of the required signals, not a sequence.*

Principle 1 (Locus). *The boundary is learned (weights) or supplied (a gate), never sampled. No post-processing of fixed weights, sampler, or preference-tuning on self-pairs yields a decision boundary.*

C7 and C8 are dual—information is what a policy can learn, retention is what it can keep—and C2 is C7’s maximum-information case.

5 Structural limits

Theorem 1 (Convergence under the axiom). *Under C1–C6, behavior cloning of π^* converges to π^* ; hence $\lim_n \text{prec}(\hat{\pi}_n) = \lim_n \text{rec}(\hat{\pi}_n) = 1$ with false abstention $\rightarrow 0$ and finite-sample rate $O(\sqrt{\text{VC}(\mathcal{H})/n})$, where $\text{VC}(\mathcal{H})$ is the Vapnik–Chervonenkis dimension of the hypothesis class.*

Proof sketch. With degenerate target $T = \delta_{\pi^*}$, the population cross-entropy minimizer is the conditional mode $= \pi^*(x)$ almost everywhere; C1 places $\pi^* \in \mathcal{H}$, C5—empirical risk minimization over $\mathcal{H}$ under T —makes it the ERM limit, C3 gives support almost everywhere, C4 and C6 make the label 1_A exact. For the finite-sample rate: by Hoeffding’s inequality union-bounded over the growth function of $\mathcal{H}$, with probability at least $1 - \delta$, $\sup_{h \in \mathcal{H}} |\hat{L}_n(h) - L(h)| \leq \sqrt{\frac{1}{2n} (\text{VC}(\mathcal{H}) \ln \frac{2en}{\sqrt{\text{VC}(\mathcal{H})}} + \ln \frac{2}{\delta})} =: \varepsilon_n$; the ERM step transfers the excess risk of $\hat{\pi}_n$ to at most $2\varepsilon_n$, and a per-example classification error rate ε bounds claim-mismatch $\mathbb{P}[\hat{\pi}_n \neq \pi^*] \leq \varepsilon_n$ by the regret transfer inequality, so both mismatch terms vanish as $n \rightarrow \infty$ at rate $O(\sqrt{\text{VC}(\mathcal{H})/n})$ (logarithmic factors suppressed). Hence claim $= 1_A$ and cert $= 1$ on A . $\square$

Each condition’s failure has a characterized consequence.

Proposition 2 (Imitation is a fixed point; violates C2). *With $T = \pi_0$ (the base model’s own rollouts), the cross-entropy minimizer is π_0 . The limit of self-imitation is $(\text{prec}(\pi_0), \text{rec}(\pi_0))$: no calibration is gained.*

Capacity rises while calibration can fall: supervised fine-tuning (SFT) on the wrong target did not fail for lack of capability, it optimized the wrong target. This is the differential prediction that turns the axiom into science—both movements of the contrast in §?? were registered before the run.

Theorem 2 (Chow frontier ceiling; violates C4). *Let $\pi_A(x) = P(1_A \mid X = x)$. The optimal claim policy is Chow’s rule: claim iff $\pi_A(x) \geq \tau$, achieving*

$$\text{prec}^*(\tau) = \frac{\mathbb{E}[\pi_A \cdot \mathbf{1}\{\text{claim}_\tau\}]}{\mathbb{E}[\mathbf{1}\{\text{claim}_\tau\}]}, \quad \text{rec}^*(\tau) = \frac{\mathbb{E}[\pi_A \cdot \mathbf{1}\{\text{claim}_\tau\}]}{P(A)}.$$

This frontier attains $(1, 1)$ iff $\pi_A \in \{0, 1\}$ a.e.—i.e. exact Ω -certifiability requires decidability to be a deterministic function of the observed input. Otherwise $(\text{prec}^*, \text{rec}^*)$, not $(1, 1)$, is the irreducible ceiling.

Proposition 3 (Rice wall off-fragment; C4c). *Predicting Ω ’s verdict from arbitrary program text is itself a semantic property of programs and is undecidable. No proposer or head can be Ω -certifiable over all programs; the decidable fragment is not an engineering shortcut but the only territory where the boundary can exist.*

Three further structural facts bound growth. The family of Ω -certifiable term algebras is countable and forms a complexity-stratified hierarchy—strict by the time-hierarchy theorem, so the ladder must grow and no rung is terminal—and it is a directed acyclic poset rather than a chain: fragments need not nest (exact decimals and three-valued null logic are incomparable), some pairs have no certifiable join, so certifiability is measured per algebra and limits exist only along directed extensions of the oracle. Second, no total, dilution-resistant scalar summarizes a verification: this was not argued but *enumerated*—a machine-checked search over 924 exhaustive and 80,000 sampled candidate summary functions returned zero survivors, so a single trust score is a category error and Definition ?? is a pair by necessity, not taste. Third, the three classical walls compose into one statement:

Abstention off-fragment is Rice-forced; the restriction to finite, witness-checkable discovery is halting-forced; external grounding is Gödel-forced; and the frontier’s monotone growth by discovery is the Σ_1^0/Π_1^0 structure of decidability itself.

The theory is not merely consistent with Rice, Church, and Gödel— it is derived by them, which is why the limits of §?? are stated as theorems rather than as road-map items.

6 Experimental design

Oracle and benchmark. All runs share one pinned oracle (a certification authority over an exact-rational claim fragment; fingerprint recorded per run) and one frozen benchmark of contract-verification tasks in a 2×2 evaluation matrix: claims $\times$ decidable (Suite *A*) and claims $\times$ undecidable (Suite *C*). Metrics: *CertYield* (claim $\wedge$ certify on *A*), *Overclaim* (claim on *C*), *FalseAbstention* (abstain on *A*), *AbstentionRecall* (abstain on *C*). The four cells are reported together; a method that trades one axis for another is visible by construction. Model: a 1.5B code instruct model, $K = 4$ samples, greedy evaluation protocol, Wilson score 95% intervals on every rate, 0 FAULTs (the alarm raised on definite checker disagreement). Frozen originals are not citable; replicated r-* runs are.

Pre-registration discipline. Each experiment registers its decision rule before the run: the shuffled placebo’s falsifier (an *A*-yield lift would falsify C2), the gate sweep’s Pareto reading, the information-intervention null band, and the reward-fidelity prediction. The complete registration ledger—every test, the claim it challenged, and its outcome—is kept as an appendix battery (Appendix ??); the main text reports the outcomes those tests established. Throughout §??, subsections labeled *primary* carry pre-registered decision rules whose outcomes confirm or falsify stated predictions; subsections labeled *exploratory* were specified before their own runs but are reported as descriptive evidence, not as tests of the axiom.

Table 2: Controlled contrast (frozen 59-task held-out split, $K = 4$, 1.5B model). Only the training target varies. Both movements were registered before the run; the corrective arm replicated at 98.3% yield.

	base	self-SFT ($T = \pi_0$)	corrective ($T = \delta_{\pi^*}$)
Suite <i>A</i> CertYield@1 (%)	38.6	79.5	95.5
Suite <i>A</i> Overclaim (%)	30.7	14.2	4.5
Suite <i>A</i> FalseAbstention (%)	7.4	0.0	0.0
Suite <i>C</i> Overclaim (%)	58.3	86.7	30.0
Suite <i>C</i> AbstentionRecall (%)	8.3	3.3	70.0

Table 3: Teaching-set selection (oracle targets, 250 iterations, seed 0).

Teaching set	Suite <i>A</i> yield (%)	Suite <i>A</i> overclaim (%)	Suite <i>C</i> overclaim (%)	Suite <i>C</i> recall (%)
base	38.6	30.7	58.3	8.3
full (91)	95.5	4.5	30.0	70.0
random (45)	92.0	8.0	43.3	55.0
disagree (44)	81.2	13.1	13.3	85.0

7 Results

7.1 The controlled contrast on the teaching target (primary)

Same base, split, prompt, and oracle; only the target changes (Table ??). Direction predicted in advance: self-imitation raises capacity and *lowers* calibration; the corrective, oracle-derived target moves both axes toward (1, 1).

Shuffled-target placebo. To kill the alternative explanation that corrective data is merely cleaner, we destroyed prompt–target correspondence while keeping the oracle targets themselves. *A* CertYield collapsed to 1.7% (replicated: 2.8%), below base on both suites. The hypothesis that correspondence—not data quality—carries the gain survives; the cleaner-data hypothesis does not.

7.2 Information nulls (C7) and retention (C8) (primary)

Five-seed runs on a 66-task benchmark, seed band *C*-Overclaim [0, 7.14], *C*-Recall [91.67, 100], *A*-side identical every seed. Two information-preserving interventions—rewording 84/99 prompts (mechanics preserved) and reordering the teaching set—landed inside the band (97.8% yield, 0.0–2.4% overclaim): registered as null, observed as null. Retention: the joint union preserves both axes (*A* 96.7 ± 1.6 , *C*-Recall 95.2 ± 6.7 , overclaim 4.8 ± 6.7 over two seeds) where an equal-budget sequential curriculum leaves *C*-Recall at 0.0%—the C8 contrast in one line; the control-arm details and the registered dose-response sub-test are recorded with the battery (Appendix ??).

7.3 Machine teaching: the disagreement set (exploratory)

All-oracle targets, equal budget (Table ??). The 44-example disagreement-only set—tasks where the base model’s majority behaviour disagrees with the oracle-consistent policy—beats both the size-matched random set and the *full* set on the mandate axis, because the full set’s decidable-heavy majority dilutes the abstention signal. And no set dominates both axes: the full set scores higher on the decidable side (95.5) while the disagreement set scores higher on the undecidable side (13.3/85.0)—teaching-set size and selection trade the two axes against each other, the frontier reappearing in teaching-set space (an information-efficiency reading: such curves have no dominating point).

7.4 The supplied boundary: a runtime gate (primary)

A decode-time gate claiming iff $\hat{\pi}_A(x) \geq \tau$ supplies C3 for *any* proposer (Table ??). The estimate $\hat{\pi}_A(x)$ is computed by an auxiliary boundary classifier—a logistic head over frozen-encoder embeddings of the contract text (the representation studied in §??), trained once on boundary labels, *not*

Table 4: Runtime abstention gate, corrective $\circ$ gate. Pareto point $\tau = 0.8$.

	ungated	$\tau=0.5$	$\tau=0.8$	$\tau=0.9$
Coverage (%)	83	80	78	46
Precision (%)	86	89	91	100
Suite <i>C</i> Overclaim (%)	30	20	13	0
Suite <i>C</i> AbstentionRecall (%)	70	80	87	100
Suite <i>A</i> FalseAbstention (%)	0	0	0	39

by token-level logits—so the gate consumes one fixed, auditable score per input, independent of sampling. At $\tau \geq 0.9$ every gated proposer—base, self-SFT, corrective—reaches Suite *C*-Overclaim 0% with Suite *C*-Recall 100%; the frontier needs capability $\circ$ calibration, and the gate supplies the calibration factor mechanically. This is the locus principle made operational: a supplied boundary dominates every sampler under evaluation.

7.5 Visibility: the boundary is in the representation (exploratory)

A boundary model on lexical features of the contract text reaches 83.6% held-out accuracy (area under the ROC curve, AUC, 0.92); re-fit on frozen encoder embeddings it reaches 100.0% held-out (AUC 1.00). The residual gap of C4 is not information-theoretic— $I(1_A; X)$ is saturated in the representation; what fails is the decode-time policy’s *use* of it (C4b), the exact ceiling of Theorem ?? made measurable.

7.6 Reward laundering, and the fix lives in the reward (primary)

A premise-fidelity probe—part of the registered hygiene battery—quantified how verifier-shaped rewards can be gamed: the original corrective run converted honest failure into certified-looking success on 7/52 outputs by strengthening an *undeclared* precondition (artifact-sound, task-violating). Repair: score `premise_invention` = -1.0 in the reward itself, then retrain. Registered prediction, measured: inventions $7 \rightarrow 0$, *C*-Overclaim $26.5 \rightarrow 0.0\%$, *C*-Recall $67.6 \rightarrow 100.0\%$, at a cost of 2.2% *A*-side false abstention. A sound oracle is not sufficient; the reward must see the task-level obligation. The reward, not the model, is the trust boundary. (Full incident record: Appendix ??.)

7.7 Cross-backbone: capacity transfers (exploratory)

On three vendors’ backbones (compatibility-gated), oracle targets transfer *capacity* completely: a 1.7B model reaches 99.3% Suite *A*-yield at 0.0 Suite *A*-overclaim, and a 3.8B model with a union recipe reaches Suite *C*-Recall 95.2 ± 4.1 . Judgment quality at equal size is close but not identical. These runs are preliminary and *co-varied*: model capacity, architecture family, lineage, training recipe (corrective vs. union), and seed count ($n = 1$ vs. $n = 3$) all differ between arms, so the table supports the capacity-transfer claim only; disentangling judgment transfer requires a controlled factorial study, which we defer to future work (§??). The full five-arm record with standard deviations is kept with the battery (Appendix Table ??); the gap narrows with scale.

7.8 Pairing integrity as a tamper detector (exploratory)

The corrective run admits a cheap consistency check: the pairing contrast $CC = \log p(\text{prompt, mismatched target}) - \log p(\text{prompt, matched target})$. Base and self-imitation give $+0.73$ nats; corrective $+0.81$; shuffled targets $+0.007 \approx 0$. *CC* responds monotonically to shuffling dose ($0 \rightarrow +0.81$ over 0/25/50/75/100% corruption) with calibration $CC(f) \approx 0.81f$, yielding a corruption detection radius $f^* \approx 1\%$ at $n = 59$ that shrinks as $\sqrt{n}$. Pairing quality becomes a measured quantity rather than a hope—the statistical shadow of the shuffled placebo.

7.9 Coverage ceiling: where no rung can help (scope audit)

On a pytest corpus, a frontier triage of all 2,148 abstentions (1,740 distinct claims) found 2,146 runtime-state assertions, 2 oracle gaps, and 0 decidable: no exact mechanism can raise coverage because nothing is left for a rung to decide. The standing rule follows: state the scope limit; never

quote coverage over an unaddressable population; instance counts rank mechanisms *within* a corpus, never across. The opposite face of the same law is measured on an addressable population: there, a single witness rung (refutation by exhibited counter-box) moved instance coverage from 7.9% to 86.1% in one step—rungs move the frontier $U \rightarrow A$ only, but where the population is addressable they move it far.

8 External Ground Truth: Protocol and Measurement

Every number above is oracle-relative by construction: $\mathcal{A}$ is defined by our own instrument. We therefore first fix—as a formal blueprint, registered before running it—the evaluation that removes this circularity, and then report the run.

Corpus and label census. An external propositional-entailment library with published ground truth: theorem and non-theorem labels, 274 problems, 2,093 axioms, median 14 atoms per problem. Ground truth is *external*—not produced by, and not editable through, the system under test. The census is not uniform and the protocol does not pretend it is: 143 problems carry labels a *classical* procedure may score (142 theorem-side—classical Theorem or intuitionistic Theorem, since intuitionistic validity implies classical validity—and 1 classical CounterSatisfiable); the 109 intuitionistic Non-Theorem labels are unusable for a classical procedure (the reading that once produced 42 false certificates) and are scored separately below by an intuitionistic engine.

Decision procedure. Propositional entailment is decidable, so the authority is a total decision procedure with checkable certificates: bitset truth-table enumeration under a budget-as-table policy ($\text{max_atoms} = 16$, 65,536 assignments—the median-14 population this protocol names; at this size a problem is one bitset pass). Every theorem label admits a witness; the non-theorem label admits a counter-assignment. Every certificate is re-derived by the detached kernel before it counts, under strict discharge: a claiming obligation must be re-derived, not merely not-contradicted. A certificate whose expanded obligation string would exceed 64 KiB is never built—the expansion of a biconditional chain is exponential in paths, not nodes—so the policy abstains rather than truncates. The (1, 1) point is therefore constructible on this corpus within the disclosed budget: exact yield and exact abstention are achievable, and the pair measures the learner rather than the instrument.

Protocol (registered in advance). (1) Fix the corpus and split; the labels never move. (2) Train or gate the learner against decision-procedure verdicts only. (3) Report (prec, rec) on the held-out theorems and non-theorems with Wilson intervals, together with the authority fingerprint. (4) Falsifier: any measured pair strictly below (1, 1) is attributable to the learner, since the instrument is exact by construction; conversely, (1, 1) here certifies transfer of the teaching recipe to an external ground truth. (5) Compare mechanisms by their *instance value*: the same rung may be worth zero on the 33,000 vendored test units of §?? (no external ground truth) and worth 274 on this corpus—mechanism counts are meaningful only relative to a declared reference class.

Result: the classical engine. Steps (1) and (3) have been executed. On the addressable labeled population the pair is *measured* at (1, 1): 89 of 89 claims correct, precision = recall = 1.000, Wilson 95% intervals [0.959, 1.000] on both axes, zero false certificates against the labels, zero kernel alarms, one authority digest (6f12977bd91dc3e0). Over the full labeled population the pair is (1.000, 0.622) with recall interval [0.541, 0.698]; the 54 misses are disclosed abstentions (47 problems above the 16-atom budget, 7 certificates above the fit limit), not errors. The falsifier (step 4) returns empty on the instrument layer.

Cross-validation against third-party artifacts. Circularity is not discharged by assertion. Three independent checks: (i) a second pipeline—a fresh recursive-descent TPTP parser with no shared parser code, Tseitin CNF, and an industrial SAT solver—agrees with the engine on every co-decided problem (zero mismatches) and confirms 141 of the 142 theorem labels with zero falsifiers (the one remainder is the second engine’s own disclosed wall-clock abstention); (ii) ILTP’s published corpus classification (“exactly one classically invalid problem”) matches the second engine’s classical-invalid set in both count and identity (SYN916+1); (iii) the header extraction agrees on 235 of 235 decided problems with the published five-prover comparison tables (zero disagreements). All 47

budget abstentions on labels were audited by the unbounded pipeline: 47 labels confirmed, zero contradictions—the abstentions are coverage, not hidden error.

Extension: the intuitionistic axis. Intuitionistic Non-Theorem does not imply classical non-theorem, so the 109 non-theorem labels require an intuitionistic engine: memoized LJ proof search (claims carry proof trees, re-checked rule-by-rule by a detached checker) and Kripke countermodel search (claims carry explicit frames, re-checked for order, monotonicity, and forcing). Registered budgets: 250,000 sequent visits; worlds ≤ 4 with per-world atom gates $\{1:20, 2:10, 3:6, 4:5\}$; 4M valuations. On the 237 intuitionistically labeled problems:

$$(\text{prec}, \text{rec}) = (1.000 [0.958, 1.000], 0.371 [0.312, 0.434]),$$

88 claims—37 checked proofs and 51 checked countermodels (49 two-world, 1 three-world, 1 one-world)—all 88 correct against the labels, zero false certificates, zero checker rejections. The 149 misses decompose into disclosed budgets and search strategy: 91 proof-visit budget, 55 no-result, 3 model budget. The precision axis is clean in both logics; the coverage axis is honest about what the engine cannot yet find. Sanity gates were registered before the corpus run—excluded middle, double-negation elimination, Peirce—each required to refuse; the double-negation gate caught a real unsoundness first: NNF negation-collapsing made $\neg\neg A \rightarrow A$ provable, which is classical reasoning, not intuitionistic.

What remains open. Step (2)—training or gating a *learner* against decision-procedure verdicts on this corpus—has not run: these numbers score the instrument and its disclosed claim policy, not a learned proposer, so the transfer half of step (4) stays open. The classical non-theorem axis rests on $n = 1$ external label, and the intuitionistic recall is bounded by a deliberately naive proof search (no focusing) and by frame/atom gates, both disclosed.

9 Limitations

The C4 wall is a theorem: exact $(1, 1)$ requires the boundary to be a deterministic function of the input; where it is not, the Chow frontier is the correct characterization of achievable performance. Off-fragment, Rice bounds all proposers. Coverage ceilings mean reported recall is conditional on an addressable population; on the external corpus the addressable labeled population is 89 of 143 at the disclosed budget, and the intuitionistic engine’s recall (0.371) is bounded by naive proof search and frame gates. Intuitionistic certificates are checked by code that shares their parser (the parser axis is differentially tested against a second independent implementation, not eliminated); one of the 142 theorem labels remains unconfirmed only because the second engine’s own wall-clock budget expired on it. A remains oracle-circular outside the external instance. Cross-backbone calibration rests on $n = 3$ with an intact confound; several cells are single-seed; hardware nondeterminism means fixed seeds fix intent, not bits. The teaching experiments run at one capacity point (1.5B, rank-8 adapters, ~ 1.6 epochs); the structural claims are scoped there. Finally, certificates certify claims *about a stated artifact*, never that the artifact is faithful to the world.

10 Conclusion

Ω -certifiability makes certified ignorance measurable as a pair, and the axiom makes reaching it falsifiable; the experiments bear the program out: oracle-derived targets cause calibration (controlled contrast $38.6 \rightarrow 95.5\%$ yield, replicated at 98.3%, with the shuffled placebo collapsing to 1.7%), information-preserving interventions are null (five registered, five observed), union preserves what sequence destroys, a supplied gate drives overclaim to 0% for every proposer, and a fidelity-aware reward closed a measured laundering channel ($7 \rightarrow 0$); and the external protocol now reports a measured $(1, 1)$ with zero false certificates in two logics under three-layer cross-validation. The limits are structural theoretical bounds—Chow, Rice, the zero-of-2,148 ceiling—not engineering artifacts, which is what makes the surviving results trustworthy: every number above was earned inside a boundary the theory drew in advance.

11 Future Research

This work is one component of research on hybrid neurosymbolic systems conducted at **Optimality.ai**, in which neural models propose and symbolic authorities dispose under a fail-closed contract; Ω -certifiability is the measurement theory for such systems. The open technical questions are as follows.

First, finish the protocol of §??: the instrument layer is executed (measured $(1, 1)$ on the addressable external population, zero false certificates, three-layer cross-validation), but step (2) remains—train or gate a *learner* against decision-procedure verdicts on this corpus and run the transfer test of step (4); grow the intuitionistic axis by replacing the naive proof search with a focused calculus and widening the countermodel gates, each a disclosed budget. Second, consolidate the single-seed cells to the ≥ 3 -seed bar with lineage-controlled cross-backbone designs—the factorial study that disentangles capacity, architecture, and training regime, all co-varied in the preliminary transfer runs of §??. Third, extend the axiom to *composed* oracles (the multi-skill setting of companion work) and characterize when pairs compose. Fourth, grow Ω only by conservative extension, tracking frontier monotonicity as a standing meta-theorem. Fifth, keep teaching from targets the oracle itself derived, with the pairing contrast (CC) as the standing tamper check; and formalize certified ignorance as a checkable judgement $\text{UNKNOWN}_\Omega(\varphi)$ whose own disposals route through the same propose–dispose loop.

A Battery of Tests

The main text carries the results; this appendix carries the tests that bought them—every registered probe and control arm, kept as a battery rather than a narrative.

A.1 Registration ledger

Table 5: Every pre-registered test, the claim it challenged, and its outcome. Falsifiers that fired are recorded as fired.

test	challenge	outcome
controlled contrast	whether the teaching target moves both axes	38.6 $\rightarrow$ 95.5 yield, 58.3 $\rightarrow$ 30.0 overclaim — as registered
shuffled placebo (C2 falsifier)	whether correspondence, not data quality, carries the gain	yield 1.7% (rep. 2.8%) — falsifier fired, C2 survives
information interventions (C7)	whether paraphrase/reorder move the pair	inside five-seed band — null, as registered
sequence dose–response (C8 sub-test)	whether harm scales with stage dose	registered monotone; observed step — sub-test falsified, C8 step confirmed
reward fidelity (λ_p)	whether an artifact-level oracle suffices	channel 7/52 found, repair 7 $\rightarrow$ 0 confirmed
CC tamper radius	whether pairing quality is measurable	$f^* \approx 1\%$ at $n = 59$, $\propto 1/\sqrt{n}$
coverage ceiling triage	whether rungs can raise pytest coverage	0/2,148 decidable — ceiling stated
counterexample-repair (CEGIS) arm	whether counterexample feedback repairs a learner	0/39 vs 13/39 plain retry — falsified: feedback induces abstention, not repair
diversity by temperature	whether sampling diversity moves the pair	near-null — registered in the null family
degenerate-abstainer guard	whether an abstain-everywhere policy can score as calibrated	vacuous precision, zero recall rejected by construction — guard holds

A.2 Retention control arms (C8)

Equal budget, oracle targets, 66-task benchmark: the staged curriculum reaches *C*-Overclaim 69.0% with *C*-Recall 0.0% against the static full set’s 0.0% / 100%; the joint union preserves both axes

(A 96.7 ± 1.6 , C -Recall 95.2 ± 6.7 , overclaim 4.8 ± 6.7). The registered dose-response sub-test (monotone harm vs stage ratio) was falsified: the damage is a *step* from sequence, not a dose—the finding that sharpened C8 from “sequence is bad” to “the effective teaching set is the union.”

A.3 Cross-backbone record

Table 6: Cross-backbone transfer (mean $\pm$ sd over $n = 3$ seeds where trained). Capacity transfers completely; judgment residual carried here.

backbone / arm	Suite A yield (%)	Suite A overclaim (%)	Suite C overclaim (%)	Suite C recall (%)
subject 1.5B, corrective	97.8	0.0	0.0	100.0
1.7B, corrective ($n=3$)	99.3 ± 1.3	0.0	42.9 ± 16.1	56.0 ± 16.1
1.7B, base	33.3	29.4	53.6	0.0
3.8B, union ($n=3$)	79.1 ± 4.2	4.6 ± 2.5	3.2 ± 2.8	95.2 ± 4.1
3.8B, base	63.9	21.1	75.0	6.0

A.4 Reward-laundering incident record

The channel of §?? in full: an invention is an emitted assertion strengthening an *undeclared* precondition, so the artifact verifies while the task obligation is violated. Rate on the original corrective run: $7/52 = 13.5\%$. Controls: base and self-imitation never invent (0)—the channel is attributable to the reward shape, not the data. Repair as registered: `premise_invention` = -1.0 (params-only, directional check), retrain: inventions 0, C -overclaim $26.5 \rightarrow 0.0\%$, C -recall $67.6 \rightarrow 100.0\%$, A -yield cost 2.2% false abstention.

B External protocol: engines, budgets, and caught defects

Engine 1 (authority). TPTP propositional frontend (all 274 problems parse; the 84 zero-axiom problems are decided as closed formulas, sound because the corpus contains no `include()` to drop); bitset truth-table semantics over the subformula closure; budget table v2 (16/65,536). Certificates record the obligation as strings that the kernel re-parses and re-decides through an independent path; strict discharge as in the main text. A memoized expansion-length guard computes the exact certificate size before any string is built (13 corpus hits, all disclosed abstentions).

Engine 2 (differential). A separate recursive-descent parser producing NNF directly—no shared lexing or AST code—Tseitin CNF, and an industrial SAT solver (`python-sat/glucose4`; a hand-written DPLL with a 2M-call budget as fallback), each problem in an isolated child process under a 4 s wall budget that is *killed*, not awaited. Results: 157/157 agreement with Engine 1 at the 16-atom budget; classical- invalid set exactly $\{\text{SYN916+1}\}$; the label audit against the published five-prover tables 235/235.

Intuitionistic engine. Single-succedent LJ backward search with frozenset memoization (250k-visit budget); every claimed proof is a tree re-checked by a checker that recomputes each rule instance from the parent sequent alone; countermodels are enumerated partial orders with per-atom upsets, re-checked for reflexivity, transitivity, antisymmetry, upset-closure, and forcing (implication universally over successors). Budgets: worlds ≤ 4 , per-world atom gates, 4M valuations.

Defects caught by the apparatus before scoring. (1) The kernel’s witness guard read the empty valuation of a 0-atom refutation (SYN916+1) as “refutation without a witness”; strict discharge raised the alarm and the guard now carves out $w = \{\}$ exactly when `atoms = 0`. (2) NNF negation-collapsing made $\neg\neg A \rightarrow A$ provable in the intuitionistic engine—double-negation elimination is classical—caught by the registered sanity gates before any corpus number was produced; negation now parses as $A \rightarrow \perp$ in IPC mode. (3) On proof-search budget exhaustion the countermodel search was being skipped (43 blind abstentions in the first run); the flow now falls through, and the first run’s numbers were discarded rather than reported. These three are why the falsifier list in §?? is empty: the checkers had real work to do.

Reproduction. `run_iltp.py`, `run_differential.py`, `run_ipc_frontier.py` in `mjolnir_research/mjolnir_experiments/oracle_non_circularity/`, each carrying its pre-registration as a docstring; raw per-problem rows in the sibling `results/` JSON files.

Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction state the theory contributions, the causal claims, and the scope limits (one capacity point, $n=3$ cross-backbone) explicitly.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: A dedicated limitations section covers the Chow/Rice ceilings, oracle circularity, single-seed cells, and scope of certificate claims.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Theorems 1 and 2 and Propositions are stated with assumptions (C1-C6); proof sketches appear in the main text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Frozen benchmark, pinned authority fingerprints, seed counts, replication classes (frozen vs r-*), and pre-registered decision rules are all disclosed (Sections 5-6).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- Reproducibility requires a reasonable avenue for the reported results, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code and data release is planned with the final version; at submission the pinned protocol, fingerprints, and tables fully describe the experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 5 specifies model, K, evaluation protocol, splits, arms, and metrics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: Wilson 95

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: All runs are LoRA fine-tunes of a 1.5B model and oracle evaluation on CPU/GPU laptops; per-run costs are small and described in the artifacts.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the code of ethics appropriate to the authors' field and institutions?

Answer: [\[Yes\]](#)

Justification: The authors have reviewed the applicable code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed a code of ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Honest abstention and auditable refusal reduce overconfident automation; the paper discusses failure modes (reward laundering) and their mitigation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of machine learning).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No high-risk model or dataset is released; trained adapters are small research artifacts with evaluation gates.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Benchmark and base model licenses are standard research licenses (Apache/MIT family) and will be listed with the release.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The benchmark, evaluation harness, and ledgers are new assets and will be documented with the release artifacts.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or human subjects; all labels are oracle-derived.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- Under prevailing codes of ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human-subject research; no IRB required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the applicable code of ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [\[Yes\]](#)

Justification: LLMs are the object of study (proposer policies); their role, size, and training procedures are fully specified.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.