
Constructibility Under Test: The Ω Pair as a Predictor of Operational Failure

Eduardo Perez Denadai
Optimality.ai

Abstract

Ω -certifiability makes trust measurable as a pair: the (precision, recall) of a policy’s claim/refuse behavior against a pinned formal authority. Deployment, meanwhile, is scored with scalars—ECE, area under the risk–coverage curve, refusal correlations—and the companion theory paper argues, by enumeration, that no scalar summarization of the pair can survive degenerate policies. This paper asks the empirical question that argument owes the field: *does the pair earn its keep as measurement?* We pre-register a head-to-head over 150 operating points (25 trained runs $\times$ a gate sweep, plus always-claim and never-claim degenerates) on a frozen benchmark with a pinned oracle, splitting tasks into calibration and held-out halves. The calibration-side pair predicts held-out operational failure—the rate at which a policy routes claims on ground the oracle has not settled—with $\rho = -0.571$ (bootstrap 95% CI $[-0.673, -0.153]$, sign-stable in 400/400 resamples), ahead of ECE (+0.498) and AURC (+0.458). The Refusal Index returns the wrong sign with 55 of 150 cells undefined, and the pre-registered naive scalarization of the pair, $(1 - \text{prec}) + (1 - \text{rec})$, is falsified— $\rho = +0.021$, CI straddling zero—exactly as the theory’s no-dilution-scalar result predicts. The never-claim policy is ranked *safest* by the operational scalar itself while every scalar metric returns null; the pair reports $(\perp, 0)$, never $(1, 1)$. A swap to the model’s own reward signal as confidence preserves the ordering, with AURC within 0.016 of precision—a margin we report rather than round away. The pair’s case is structural, not a point-estimate sweep: two axes that must be read together, and vacuous precision that can never count as one.

1 Introduction

The companion paper defines Ω -certifiability: a policy π over an input space $\mathcal{X}$ is measured by the pair (prec, rec) of its claims against a pinned, deterministic authority Ω , where $(1, 1)$ means the policy claims exactly the decidable tasks and certifies every claim it makes [Denadai, 2026a]. The theory establishes when a learner can reach that point, when it cannot (Chow, Rice, No Free Lunch), and why a single number cannot summarize the pair: a scalar dilutes one axis into the other, and the degenerate policies—answer-everything, refuse-everything—break every summary the field currently uses.

Those are arguments. This paper is the measurement question they owe: if the pair is the right trust metric, three things should be demonstrable on live policies.

- (1) **Predictive validity.** Computed on a calibration half, the pair should predict *operational failure* on a held-out half—at least as well as the field’s scalars (ECE, AURC, Refusal Index) do.
- (2) **Degenerate separation.** The policies that make every scalar lie or go null must be reported by the pair as the failures they are: vacuous precision is $\perp$, never 1.

- (3) **Robustness.** The result should survive (i) resampling (bootstrap intervals on every correlation) and (ii) swapping the confidence source (external boundary head $\rightarrow$ the model’s own reward signal).

Each claim is registered before the run—including two expected *falsifiers*: we predict that the naive distance $(1 - \text{prec}) + (1 - \text{rec})$ fails as a scalar summary (the no-dilution result), and we give the Refusal Index its expected sign honestly so that its failure, if it fails, is visible as a falsification rather than explained away.

Contributions. (1) A pre-registered head-to-head protocol (§3): 150 policies, a held-out failure target, and four metrics with declared signs. (2) The primary result (§4): precision leads every scalar on the external head, with bootstrap CIs excluding zero and full sign stability; the naive pair-sum and the Refusal Index are falsified as registered. (3) The degenerate table (§4.2): every scalar metric is null on the never-claim policy that the operational scalar ranks safest—the pair reports $(\perp, 0)$. (4) A confidence-head swap (§4.4) that preserves the ordering while shrinking the precision–AURC margin to 0.016, reported at full honesty.

2 Related work

Selective prediction and the reject option. Chow’s reject option [Chow, 1957, 1970] predates the modern setting by five decades: abstain when conditional risk exceeds a threshold. Modern treatments give risk–coverage curves and coverage-optimal strategies [El-Yaniv and Wiener, 2010, Pietraszek, 2005, Geifman and El-Yaniv, 2017, Franc and Benbouzid, 2023]; AURC, the metric we use, is due to Geifman and El-Yaniv [2017]. Our operational failure target is a coverage-conditional risk; the pair differs by refusing to integrate the curve into one number and by scoring the abstain axis positively against ground truth rather than treating abstention as unobserved.

Calibration and risk control. ECE after temperature scaling is the standard scalar [Guo et al., 2017]. Statistical layers convert arbitrary scores into finite-sample guarantees: conformal risk control and its relatives [Angelopoulos and Bates, 2023], and—closest to our gate—confidence-interval calibration for risk-controlled selective answering [Dong and Shinnou, 2026]. These control the *risk among accepted answers*; none scores whether abstentions were correct decisions, which is the recall axis here, and none is defined on policies that claim nothing.

Measuring refusal. The Refusal Index defines knowledge-aware refusal as the Spearman correlation between refusal probability and error probability [Pan et al., 2026]; RefusalBench shows empirically that frontier models fail refusal and that scale does not fix it [Muhammed et al., 2026]. We evaluate RI exactly as specified (two-pass-style probabilities from K samples) and report its behavior under policies it was not designed for.

Aligning to selective prediction. RLSR optimizes AURC directly with verifiable rewards [Luo et al., 2026]; the certified-learning line optimizes claim/refuse behavior against formal authorities [Denadai, 2026a,b]. The present work sits on the evaluation side: what to *measure* once policies are trained.

3 Protocol (registered before the run)

Corpus and policies. A frozen contract-verification benchmark with a pinned oracle: every task carries an external label (decidable / undecidable) fixed before any policy runs. We take the 66 evaluation tasks common to 25 curated runs—base, corrective, self-SFT, shuffled controls, teaching variants, ensembles, and three backbones—each with $K = 4$ samples per task. Policies are the run’s first-sample behavior gated by the boundary head $\hat{\pi}_A(x)$ at thresholds $\tau \in \{\text{none}, .5, .7, .8, .9, .95\}$: $25 \times 6 = 150$ real operating points. Two degenerates are registered alongside: ALWAYS (claims every task, certifies none) and NEVER (claims nothing).

Split. Tasks are stratified by suite and split half calibration / half test (seed 0): 32 calibration (22 decidable), 34 test (11 undecidable). All metrics are computed on the calibration half only; the target is measured on the test half only.

Metric	Definition (as run)	Expected sign vs F
Ω -precision	$\sum \text{cert} / \sum \text{claim}; \text{no claims} \rightarrow \perp$ (never 1)	—
Ω -recall	$\sum (\text{cert} \wedge 1_A) / A $	+
Ω -distance	$(1 - \text{prec}) + (1 - \text{rec})$	—*
ECE	10-bin, claim-conditioned, $\text{conf} = \hat{\pi}_A$, $\text{label} = 1[\text{dec}]$	+
AURC	claims by $\hat{\pi}_A$; $\text{error} = \text{claim} \wedge \neg \text{cert}$	+
Refusal Index	Spearman(refuse-prob , error-prob), $K=4$	—

* registered falsifier: expected to fail.

Table 1: Registered metric definitions and expected signs against held-out F . All six signs were written before any outcome was computed.

Metric	Point ρ	95% CI ($B=400$)	$P(\text{reg. sign})$	n
Ω-precision	−0.571	[−0.673, −0.153]	1.00	146
ECE	+0.498	[+0.234, +0.735]	1.00	146
AURC	+0.458	[+0.293, +0.677]	1.00	146
Ω -recall	+0.302	[+0.157, +0.436]	1.00	150
Ω -distance	+0.021	[−0.145, +0.312]	0.33	146
Refusal Index	+0.188	[−0.043, +0.448]	0.06	95

Table 2: Primary head-to-head: Spearman correlation of calibration-side metrics against held-out operational failure, with bootstrap intervals and sign stability against the registered signs (Table 1).

Target: operational failure.

$$F = \frac{\sum \text{claim} \wedge \neg \text{certified} \wedge \neg \text{decidable}}{|\text{undecidable}|} \quad \text{on test,} \quad (1)$$

the Suite-C overclaim rate: the system routed a claim on ground the oracle has not settled and could not certify it—the same failure class as executing an action whose precondition was never stated.

Metrics (calibration half). Table 1 gives the registered definitions and expected signs against F . ECE is 10-bin, claim-conditioned, confidence $\hat{\pi}_A$, label 1[decidable]. AURC orders the claimed set by $\hat{\pi}_A$ with error $\text{claim} \wedge \neg \text{certified}$. The Refusal Index is the Spearman correlation of per-task refusal probability and error probability over the $K = 4$ samples, with the gate turning $\hat{\pi}_A < \tau$ into refusal 1 / error 0.

Confidence heads. Head 1 is the external lexical boundary head (features fit on train, scored on eval). Head 2 (§4.4) replaces only the metrics’ confidence with the run’s own mean reward calibration mapped to $[0, 1]$; the gate machinery is unchanged.

Resampling. $B = 400$ stratified bootstrap resamples of calibration and test (seed 0); percentile CIs on every ρ , plus the sign-stability probability against the registered sign.

4 Results

4.1 Primary: calibration-side metrics vs held-out failure

Table 2 gives $\rho(\text{metric}_{\text{cal}}, F_{\text{test}})$ across the 150 real policies.

Precision is the strongest single predictor and the only one whose CI excludes zero *in the registered direction with full sign stability*. ECE and AURC carry real signal with their expected signs—the field’s scalars measure a real thing on real policies. Recall moves *with* failure (+0.302): gating suppresses claims on both sides of the boundary together, which is precisely why the axes must be read as a pair.

The two registered falsifiers hold. Ω -distance—the naive sum of the axes—collapses to $\rho \approx 0$ with a CI straddling zero and sign stability 0.33: under gating, recall falls with F and cancels precision’s signal when the two are added. This is the no-dilution-scalar result observed on live data [Denadai,

Policy	F	prec	rec	dist	ECE	AURC	RI
ALWAYS	1.00	0.00	0.00	2.00	0.102	1.00	null
NEVER	0.00	$\perp$	0.00	null	null	null	null

Table 3: Degenerate policies. The operational scalar alone ranks NEVER the safest possible system ($F = 0$); every scalar metric is null on it. The pair reports $(\perp, 0)$: visibly not $(1, 1)$, visibly not safe-to-prefer.

Metric	Head 1 (boundary $\hat{\pi}_A$)	Head 2 (reward)
ECE	+0.498	+0.480
AURC	+0.458	+0.555
Ω -precision	−0.571	−0.571 (head-invariant)

Table 4: Second confidence head: the model’s own mean reward calibration replaces the external boundary head in ECE/AURC only; the gate is unchanged. Ordering survives; the precision–AURC margin on head 2 is 0.016, inside this experiment’s resolution.

2026a]: the pair must be read as two axes, not added. The Refusal Index returns the *wrong* sign (+0.188 against a registered $-$, sign stability 0.06) and is undefined on 55 of 150 policies—constant refusal vectors at the gate extremes make Spearman undefined. Both outcomes were registered in advance; neither is reinterpreted.

4.2 The degenerate table

Table 3 is the case the vacuous-precision rule was built for. The operational scalar ranks NEVER the *safest possible* system. Every scalar metric is null on it—there are no accepted answers to calibrate, no risk–coverage curve to integrate, no variance in refusal to correlate. The pair reports $(\perp, 0)$: precision is explicitly undefined-because-vacuous and recall is zero, so the policy is visibly not $(1, 1)$ and visibly not safe-to-prefer. Vacuous precision never counts as 1, by construction.

4.3 What the falsifiers buy

The distance falsifier is the sharpest result in the table: the companion theory proves by enumeration that no scalar summary of the pair survives all policies [Denadai, 2026a]; we registered the most natural candidate scalar, watched it fail (0.021 vs -0.571), and the bootstrap confirms the failure is not resampling noise ($P = 0.33$ for the registered sign). A metric that requires two numbers is not a stylistic preference—it is what the data demanded.

4.4 Second confidence head

Replacing the external boundary head with each run’s own mean reward calibration (Table 4): the ordering survives—precision remains the most negative ρ , and the scalar cluster stays in 0.48–0.56 on both heads. We state the margin plainly: on head 2, AURC (0.555) is within 0.016 of precision (0.571)—at this sample size that is a tie, not a win. What does not depend on the head at all is the structural result: precision never uses a confidence, the degenerate table does not change with the head, and no scalar becomes defined on NEVER.

5 Discussion

The pair’s case is structural. On point estimates, the pair’s advantage over the best scalar is thin and head-dependent; a paper that claimed a sweep would be overclaiming its own metric. The defensible claims are the registered ones: (i) precision is never *worse* than the best scalar on either head and is clearly better on the primary head with CIs excluding zero; (ii) the two-axis reading is load-bearing—the registered scalarization of the pair is falsified; (iii) the degenerate policies that every scalar scores null on are reported, with vacuity visible as $\perp$ rather than as 1 or as silence.

Reading scalars honestly. ECE and AURC work on real operating points. Their failure modes are specific: undefined on policies without accepted answers, blind to the abstain axis (NEVER’s refusals are never scored as correct decisions), and—like all scalars—silent about *which* axis moved. The Refusal Index, evaluated as specified, was the weakest instrument here: wrong sign, 37% null cells.

Position. Relative to the selective-prediction literature (§2), this paper contributes the *held-out failure target with an external label* (overclaim against a pinned oracle rather than self-reported confidence), the degenerate protocol, and registered falsifiers. Relative to the companion theory paper [Denadai, 2026a], it converts “no dilution-resistant scalar” from an enumeration argument into an observed failure on live policies.

6 Limitations

The split is small: 32 calibration / 34 test tasks (11 undecidable on test); correlations at ± 0.05 are not resolved, and the head-2 margin (0.016) is inside that noise. All policies come from one benchmark family (contract-verification tasks with a pinned oracle); F is oracle-relative overclaim, not a production incident log—the money-path corpora lack the per-instance confidence and K -sample structure this protocol needs, and porting to action logs with real incident labels is the stated next run. One confidence family (lexical boundary head, reward head); embedding heads and verbalized confidence are untested. RI’s null cells follow from constant refusal vectors under deterministic gating; a sampling-based refusal probability would change its n and may change its sign—reported as run, not as universal. Finally, nothing here tests the *training-side* axioms of the companion paper: this is an evaluation study.

7 Conclusion

The Ω pair was argued to be the right trust measurement; here it is tested as one. Against held-out operational failure on 150 gated policies, precision leads the field’s scalars on the primary head, ties the best on a second head, and is the only instrument that survives the degenerate policies with its meaning intact; its naive scalarization fails exactly as the theory predicted, and the Refusal Index fails as honestly as it was registered. The pair is not a better number—it is two numbers that must be read together, plus one symbol ($\perp$) that no scalar can express.

References

- Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.
- C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957.
- C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
- Eduardo Perez Denadai. ω -certifiability: A theory of certified ignorance. Technical report, Optimality.ai, 2026a.
- Eduardo Perez Denadai. Reinforcement learning from verifiable feedback: Teaching language models to say “i don’t know”. Technical report, Optimality.ai, 2026b.
- Sijin Dong and Hiroyuki Shinnou. Uncertainty-aware abstention in large language models with provable alignment guarantees. *arXiv preprint arXiv:2607.04430*, 2026.
- Ronen El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.
- Václav Franc and Daniel Benbouzid. Cost-based abstention in classification. *Journal of Machine Learning Research*, 24(274):1–64, 2023.

- Yonatan Geifman and Ronen El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, pages 4878–4887, 2017.
- Chuan Guo, Geordie Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.
- Gaoxiang Luo, Yifan Wu, Sinian Zhang, Aryan Deshwal, and Ju Sun. Aligning language models with selective prediction. *arXiv preprint arXiv:2607.03528*, 2026. ICML 2026 AIWILD Workshop.
- Aashiq Muhamed, Leonardo F. R. Ribeiro, Markus Dreyer, Virginia Smith, and Mona T. Diab. RefusalBench: Generative evaluation of selective refusal in grounded language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 6811–6856, 2026.
- Wenbo Pan, Jie Xu, Qiguang Chen, Junhao Dong, Libo Qin, Xinfeng Li, Yu Haining, and Xiaohua Jia. Can LLMs Refuse Questions They Do Not Know? Measuring Knowledge-Aware Refusal in Factual Tasks. In *International Conference on Learning Representations (ICLR)*, 2026.
- Tadeusz Pietraszek. Optimizing abstaining classifiers using risk–coverage curves. In *Proceedings of the 14th European Conference on Machine Learning (ECML)*, pages 349–360, 2005.