The Missing "I Don't Know": Why Three Reasoning-Reliability Findings Converge on Calibrated Abstention

Anonymous ACL submission

Abstract

Three recent results describe what look like unrelated LLM reliability problems. Yin et al. (2026) show reasoning RL collapses tool-reliability representations. Suleymanov et al. (2026) show that under safety-constrained generation, large models rewrite flagged spans while small models truncate. Bastounis et al. (2024) prove any consistent-reasoning system without an implicit "I don't know" function must hallucinate infinitely often on broad problem classes. We argue these findings converge on a single intervention: calibrated abstention is what each independently identifies as the missing capability, even though the unavailability they document — a capability gap, a policy gap, and a recursion-theoretic gap — has a different source in each case. Honesty post-training has narrowed the gap in deployed models, but principled closure of the class Bastounis identifies requires a calibrated abstention function whose training signal at the leaderboard level is absent: dominant benchmarks assign zero reward to decline, so the leaderboard gradient that would select for the function does not exist. We propose four changes to evaluation: triple-scoring, abstention-rate reporting, capability-stratified evaluation, and mandatory calibration metrics. Benchmark reform is necessary, not sufficient, for closing the gap the theorem identifies.

1 Introduction

Three results from the past two years describe what looks like three different problems. Yin et al. (2026) show that reinforcement learning for reasoning collapses the internal representations that govern tool reliability, producing agents that invoke tools they know are unavailable. Suleymanov et al. (2026) show that when an agentic loop constrains outputs away from sensitive completions, large models rewrite affected spans into nuanced alternatives while small models truncate text to the point of non-utility. Bastounis et al. (2024) prove formally that any system that reasons consistently across paraphrased equivalents must hallucinate infinitely often on a broad class of problems, including basic arithmetic, unless it employs a sufficient escape such as the implicit "I don't know" function.

We argue these findings converge on a single intervention. The unavailability each documents has a different source — a capability gap (Yin), a policy gap (Suleymanov), and a recursion-theoretic gap that is agnostic to source (Bastounis) — but in each case the failure mode is what happens when a system without calibrated abstention is forced to commit on inputs where commitment is unwarranted. Yin and Suleymanov are not instances of Bastounis's theorem; they are independent empirical settings whose failure modes are most parsimoniously addressed by the same construction the theorem identifies. We treat the three results as convergent evidence for calibrated abstention rather than as instances of a single structural phenomenon, and argue in Section 3 that the convergence has policy consequences the field has not yet acted on.

This matters because most work on the reliability problem has treated it as a tuning problem. Better RL objectives, better prompting, better alignment. Honesty post-training has narrowed the gap in deployed models. But the theorem rules out closure in principle for any consistent-reasoning system that lacks the implicit "I don't know" function, and current training pipelines do not compute it at the structural level the theorem requires. A parallel line of abstention-specific work (Kamath et al., 2020; Wen et al., 2025; Kirichenko et al., 2025) has developed mechanisms but has not produced reform of the dominant leaderboard benchmarks, which is where the training signal is set. The gap that the theorem identifies is closeable only by a calibrated abstention function, and the training signal for that function is currently backwards.

Two claims follow. First, the three findings con-

verge on calibrated abstention as the warranted intervention, and treating them separately has obscured that convergence. Second, benchmark reform is necessary — though not sufficient — for closing the reasoning-reliability gap that the theorem identifies. Honesty post-training partially substitutes for the missing signal; principled closure requires that benchmarks stop assigning zero reward to decline. We propose triple-scoring across correct, wrong, and abstained outcomes, abstention-rate reporting, capability-stratified evaluation, and calibration metrics (ECE, Brier) alongside accuracy.

2 The Three Findings

2.1 Reasoning RL collapses tool-reliability representations

Yin et al. (2026) build SimpleToolHalluBench, which tests two failure modes where abstention is the correct behavior: queries where no tool is available and queries where only a distractor tool is present.

They find that reinforcement learning targeting reasoning raises tool hallucination in direct proportion to reward-signal gains. The effect holds even when the RL training set contains no tools: GSM8K math fine-tuning increases hallucination on tool benchmarks the model never saw during training. The effect appears both under GRPO-style outcome-centric RL and when reasoning is only elicited at inference time by prompt switching. The mechanism is a representation collapse: tool-reliability pathways destabilize in early and middle layers while general reasoning pathways stay stable, and activation probes localize the resulting errors as divergences concentrated in late-layer residual streams. Mitigations fail cleanly. Prompt instructions give marginal gains. DPO reduces hallucination only by sacrificing utility. The authors call this a reliability-capability trade-off with no known resolution.

2.2 Safety-constrained generation produces scale-dependent failure modes

Suleymanov et al. (2026) study what happens when an agentic self-correction loop (SemSIEdit) flags and rewrites spans containing sensitive inferences. Across 13 models they document the rewrite-versus-truncate divergence. GPT-5 responds with constructive expansion, increasing mean answer length by 110 characters (Suleymanov et al.'s headline figure across the GPT-5 condition) as it rewrites flagged spans into nuanced alternatives while keeping utility at 8.58/10. Smaller and mid-sized models respond with destructive truncation, reducing length by 50–73% across the smaller-model conditions reported. Both strategies reduce leakage; only the first preserves utility. The authors argue the Editor role requires a minimum reasoning threshold, below which agentic refinement collapses into refusal. The bimodality is presented at the headline level; we cite the headline numbers and do not infer significance beyond the source paper's own qualifications.

A second finding inverts the direction. Toggling chain-of-thought on Qwen3-8B with no defense raises leakage from 74% to 84% (single-run point estimates as reported): reasoning lets the model connect benign facts into sensitive inferences. Running the same model with SemSIEdit active is consistent with sign inversion in the reported numbers, dropping leakage to 42% versus 47% without reasoning (single-run point estimates). The capability that amplifies the failure mode unconstrained is the same capability that powers the rewrite under constraint.

2.3 Consistent reasoning without abstention is provably unreliable

Bastounis et al. (2024) prove the Consistent Reasoning Paradox. Consistent reasoning, in their formal sense, is the property of a Turing-complete system that returns the same answer across a computably enumerable family of paraphrased equivalents of a task ("Tell me the time!" and "What is the time?"). The theorem establishes that any such system that always answers will hallucinate infinitely often on certain problem classes, and that basic arithmetic contains such a class. A sufficient construction that escapes the paradox is an implicit "I don't know" function: a partial computable refusal that fires on the subset of inputs where consistent reasoning would force confabulation. Bastounis et al. do not claim this is the unique escape, nor that practical calibrated abstention — a softmax-thresholded refusal head, a conformal selection layer (Quach et al., 2024), or a Kadavath-style $P(\mathrm{True})$ self-evaluator (Kadavath et al., 2022) — instantiates the recursion-theoretic object the theorem describes. We adopt the I-don't-know function as a target the missing benchmark gradient should select for, while flagging that the theorem's formal object and its operational ana-

logues are not the same.

Each of the three papers stands on its own. Yin's work is a mechanistic empirical study. Suleymanov's is behavioral. Bastounis's is a theorem in computability theory. Taken separately they explain different phenomena. We argue in Section 3 that they converge on the same intervention.

3 Convergent Evidence

The three findings converge on the same intervention. Each documents what happens when a system that lacks calibrated abstention is forced to commit on inputs where commitment is unwarranted, but the source of the unwarranted commitment differs across the three. Yin documents a capability gap: the model lacks the tool, the deductive path, or the fact, and outcome-reward RL has selected against the only honest response. Suleymanov documents a policy gap: the surrounding agentic loop has flagged the natural completion as sensitive, leaving the model to produce something else in its place. Bastounis proves the recursion-theoretic case: even an idealized consistent-reasoning system must hallucinate infinitely often on inputs where the implicit "I don't know" function would fire. Three different unavailabilities, three different empirical regimes, one shared remedy: a calibrated abstention construction the theorem identifies as one sufficient escape from its impossibility result.

We use "convergent evidence" deliberately, and not "one phenomenon" or "a single structural property." The theorem applies to systems achieving recursion-theoretic consistency over computable paraphrase classes; GRPO selects on outcome reward over rollouts of the same prompt and does not, in general, achieve that property. The simpler claim suffices: outcome-reward RL with no abstention reward selects against decline across the entire training distribution, the model is therefore optimized to commit on inputs where commitment is unwarranted, and what closes that gap in principle is the I-don't-know construction the theorem identifies as one sufficient escape. The convergence is at the level of the missing capability, not at the level of the formal property each system achieves.

This frame absorbs Yin without invoking the theorem mechanistically. Reasoning RL on outcome rewards selects against partial answers and honest failures relative to coherent commitments. Running the training on GSM8K still amplifies tool hallucination because what RL is teaching is not tool behavior; it is the policy of always producing a numeric output. When the evaluation distribution shifts to queries where the correct response is decline, the policy generalizes anyway. Late-layer residual divergence is the activation-space signature of a model trained to produce numeric outputs producing them where it should not. Kirichenko et al. (2025) build an abstention-specific benchmark across 20 datasets and report that reasoning fine-tuning degrades abstention by 24% on average (a single-run mean across 20 datasets, with task-level variance Kirichenko et al. report; we cite the headline figure without inferring significance beyond it), including in domains the reasoning models were explicitly trained on. The empirical pattern is consistent with what the theorem requires: a system without the I-don't-know function commits where decline would be correct. Whether the precise mechanism is recursion-theoretic consistency-forcing or simpler selection-against-decline, the proposed remedy is the same.

This resolves an apparent tension between Kirichenko and Suleymanov. In Kirichenko's setting, expanded test-time compute worsens abstention as the model uses extra tokens to manufacture missing context. In Suleymanov's, expanded compute enables the constructive rewrite. Reasoning capability is dual-use: its impact on reliability depends on whether the surrounding loop rewards honest refusal or mandates a definitive output. The two settings differ in their loops, not in their reasoning.

Suleymanov fits the same convergence from the opposite direction and tells us what varies with capability. SemSIEdit creates regions where the model's natural completion has been flagged as sensitive. The model must produce something in those regions; what it produces depends on whether it can compute a local "I don't know, here is what I can still say" construction at each flagged span. GPT-5 can. The +110 character expansion that Suleymanov et al. (2026) report (a headline mean across the GPT-5 condition) is the empirical shape of an abstention-with-context intervention rather than a measured calibrated abstention in the technical sense. Gemma-3-4B cannot, and so cuts rather than rewrites. Suleymanov's setting is technically prescriptive refusal (the policy rules out the natural completion), distinct from Kirichenko et al. (2025)'s epistemic-abstention frame. The convergence we claim is at the gradient-level mechanism that operates in both cases — the model is required

to interrupt a probable token sequence and substitute a calibrated alternative — not at the philosophical character of the unavailability.

The same reasoning capacity that amplifies leakage when no constraint is active produces safer output when the constraint is active, because the model can now spend compute on the rewrite rather than on connecting facts into sensitive inferences. Reasoning without abstention infrastructure is what Yin documents. Reasoning with the beginnings of an abstention infrastructure is what Suleymanov's GPT-5 result looks like. Bastounis's theorem tells us why the distinction is structural, not incidental.

4 Why Current Benchmarks Make This Worse

The four benchmarks that dominate LLM leaderboards (MMLU (Hendrycks et al., 2021), HellaSwag, HumanEval, GSM8K) all share a property. None provides an abstention option with nonzero reward. Table 1 shows the structure. Across all four, expected reward for any attempt strictly exceeds reward for decline, so guessing dominates even when the model has no signal. A model optimized against these leaderboards learns the policy the scoring function rewards.

Yin's GSM8K-to-tool-hallucination result (Section 2.1) is what this incentive structure produces downstream when the benchmark is also used as an RL training signal. GRPO's group-relative advantages are computed from outcome rewards on GSM8K answers, and those rewards are zero for any response that fails to produce a numeric match. No reward signal touches "decline." The model is selected for producing numeric outputs whether or not they are correct, and the policy generalizes beyond the training distribution to evaluation tasks where decline would be the correct response. We distinguish two pressures the table reflects: *gradient pressure* (the actual RL training signal, which is what GSM8K-as-RL-data exerts) and *leaderboard pressure* (post-hoc selection on model releases via evaluation, which all four benchmarks exert). Yin's result is a downstream effect of the first; the proposal in Section 5 addresses the second, and only some of the gradient-pressure case via the same reform.

Error-rate audits compound the problem. Gema et al. (2025) find that 6.49% of MMLU questions contain ground-truth errors, rising to 57% in the Virology subset (where the small denominator widens the Wilson-interval bound considerably). A calibrated model that refuses to commit on an ambiguous or mislabeled item scores identically to a model that commits to the labeler's mistake. The benchmark cannot distinguish appropriate abstention from failure.

A model with no gradient pointing toward "I don't know" is what the three papers of Section 2 document.

5 A Concrete Proposal

Prior work has developed the mechanisms we invoke and several adjacent benchmark reforms. The selective-prediction literature (Geifman and El-Yaniv, 2017; Kamath et al., 2020; El-Yaniv and Wiener, 2010) provides the risk-coverage framework that Equation 1 below specializes to a three-action benchmark setting. Mechanism-level abstention work has produced semantic-uncertainty estimates (Kuhn et al., 2023), refusal-tuning (Zhang et al., 2024), conformal generation (Quach et al., 2024), self-knowledge calibration (Kadavath et al., 2022), and training-time ternary-reward RL (Wei et al., 2025; Zhai et al., 2026; Ralethe and Buys, 2025) that optimizes three-way rewards involving abstention directly during model training, surveyed broadly in Wen et al. (2025). Existing benchmark reforms address adjacent dimensions of the same problem: HELM (Liang et al., 2023) introduced multi-metric leaderboards including calibration; TruthfulQA (Lin et al., 2022) penalizes confident-but-wrong answers; AbstentionBench (Kirichenko et al., 2025) measures abstention directly; FActScore (Min et al., 2023) decomposes factuality. What has not been argued is that these reforms compose into a structural requirement, not an optional companion to accuracy. We ground necessity in the convergent-evidence argument of Section 3 rather than in a direct application of the theorem, since current LLMs do not strictly inhabit the theorem's formal class; the structural lesson the theorem teaches — that without an abstention construction certain failures are inevitable — generalizes, and the proposed bundle is what selects for the construction. The contribution below is the necessity claim and the specific bundle, not the individual reforms; we position each proposal against its closest existing analogue. Wei et al. (2025), Zhai et al. (2026), and Ralethe and Buys (2025) use closely related training-time three-way rewards (with action-level vs. outcome-level variants), but

Benchmark	Output format	Attempt reward	Abstain reward
Multiple-choice (random guess strictly dominates abstain):			
MMLU	4-choice MCQ	0.25 (uniform guess)	0
HellaSwag	Sentence-completion MCQ	0.25 (uniform guess)	0
Open-generation (any output rewarded over none):			
HumanEval	Unit-test pass@k	$\geq 0, > 0$ in expectation	0
GSM8K	Exact-match numeric	$\geq 0, > 0$ in expectation	0

Table 1: Expected scores under four dominant LLM benchmarks. Reward for any attempt weakly exceeds reward for decline in every row. The argument is sharper for multiple-choice (random guessing strictly dominates abstention regardless of model knowledge) than for open-generation (where the inequality follows by construction whenever the model has any signal above noise).

all operate on the training objective; our claim is that the dominant evaluation pipelines remain broken even where individual training pipelines have adopted such rewards, so without leaderboard-side reform these training-time advances stay isolated rather than compounding. We propose four changes.

1. **Triple-scoring.** Replace the binary correct/incorrect scoring function with a three-valued outcome: correct, incorrect, or abstained, where *abstain* is a distinct action the model selects rather than a missing prediction. For open-weight base models that do not emit a structured *abstain* action, the action can be derived post-hoc from predictive entropy thresholding, semantic-uncertainty clustering across multiple stochastic generations (Kuhn et al., 2023), or conformal selection layers (Quach et al., 2024); the choice does not affect the design condition below. Let $p(c \mid x), p(w \mid x), p(a \mid x) \geq 0$ with $p(c \mid x) + p(w \mid x) + p(a \mid x) = 1$ denote the model's empirical probabilities of correct, incorrect, and abstain outcomes on input x, estimated from k rollouts under the model's decoding distribution. Abstention receives positive weight on problems where the model is outside its reliable capability and zero weight elsewhere. The design condition is the reward inequality

$$R_{\text{abstain}} > p(c|x) \cdot R_{\text{correct}} + p(w|x) \cdot R_{\text{incorrect}} \tag{1}$$

where $R_{\text{correct}} > 0$ is the reward for a correct attempt and the right-hand side is the expected return conditional on the model attempting an answer. Equation 1 is the standard cost-sensitive abstention threshold in the tradition of Chow (1970) and the selective-prediction literature it grounds: the inequality fixes the model's threshold on subjective probability at $p^* = R_{\text{abstain}}/R_{\text{correct}}$ when $R_{\text{incorrect}} = 0$, and triple-scoring is therefore not a new mathematical object but a reapplication of a classical decision-theoretic rule to a leaderboard setting where the abstain action has historically been priced at zero. The choice of $R_{\text{abstain}}/R_{\text{correct}}$ is therefore a consequential leaderboard-design decision rather than a mathematical given: it sets p^* and so encodes the operational tolerance for being wrong relative to being silent. Because no single threshold fits all downstream domains — creative writing, medical QA, and legal advice impose very different costs on incorrect commitments — a useful reporting convention is to publish performance across a sweep of $R_{\text{incorrect}}$ values (or equivalently p^* values) rather than a single point estimate, analogous to a Receiver Operating Characteristic curve, so that downstream developers can read off how a model's abstention rate changes with the simulated cost of failure and select the operating point that matches their domain's risk profile. Under current scoring ($R_{\text{abstain}} = 0$, $R_{\text{incorrect}} = 0$, $R_{\text{correct}} > 0$), the inequality strictly fails for any $p(c \mid x) > 0$, which is why guessing dominates over decline.

Deployment significance. This inequality matters in deployment. In high-stakes settings such as medical QA, legal advice, or autonomous systems, the actual-use $R_{\text{incorrect}}$ is sharply negative, often far more so than R_{abstain}. Benchmarks that score both at zero erase the asymmetry. Models trained on them are then deployed into contexts where the

asymmetry is strong, with no learned signal for when to refuse. The scope of this scoring rule is epistemic abstention (the model recognizing a capability gap), not prescriptive refusal (the model executing a safety policy on a dangerous query). Safety refusal is scored separately by safety benchmarks and should not be folded into capability scoring.

Goodhart coupling. Triple-scoring on its own is Goodhart-vulnerable: a model that abstains on every hard item collects the abstention reward on the unanswerable slice while paying nothing on the answerable slice except the foregone correct reward, and blanket abstention is a competitive strategy unless calibrated commitment is rewarded above it. We therefore couple Equation 1 with a strictly proper scoring rule (Brier, 1950; Gneiting and Raftery, 2007; Savage, 1971) on the model's expressed confidence (Brier or log-score) over its committed answers: a model that commits with calibrated confidence beats a model that abstains uniformly, and a model that commits over-confidently beats neither. Without the coupling, the abstention rate is unbounded above; with it, the abstention rate converges to the rate at which the model's calibrated confidence falls below p^*. The equilibrium for a calibrated model is therefore to abstain on inputs where its expressed confidence falls below p^* and commit elsewhere — not the trivial "always abstain" that would game the bare inequality.

Ground-truth requirement. Scoring reform also requires that benchmark ground truth be clean: Gema et al. (2025) report a 6.49% mean error rate on MMLU, rising to 57% on the Virology subset (where the small denominator widens the Wilson-interval bound considerably), so a calibrated model that correctly abstains on a flawed item is penalized against the labeler's mistake under any scoring scheme. Decontamination is a prerequisite for the proposal, not an independent concern; the MMLU-Redux dataset (Gema et al., 2025) provides 5,700 manually re-annotated questions across all 57 MMLU subjects and serves as a drop-in replacement for any MMLU-based deployment of triple-scoring.

2. **Abstention rate as a primary metric.** Papers should report abstention rates alongside accuracy, broken out by problem stratum. A model at 0.85 accuracy with 5% abstention is doing something different from one at 0.85 accuracy with 40% abstention, and leaderboards currently cannot tell us which we have.

3. **Capability-stratified evaluation.** Benchmark construction should identify problem classes where abstention is the correct behavior. Two kinds of stratification are possible. Environment-level unanswerability, as in Yin et al. (2026), is a ground-truth fact about the task: "what is the current time in Park Forest Village" given no tool access has no correct completion, and the benchmark can score refusal accordingly. Model-level unanswerability, as in factual QA where the answer depends on whether the model has the fact, is harder, because what counts as "outside capability" is model-dependent. The environment-level case is tractable now and should be adopted. The model-level case is an open problem for benchmark design. Acknowledging the distinction, rather than collapsing it, is what separates a usable proposal from a wishlist.

4. **Calibration metrics as mandatory companions.** Expected calibration error (ECE) and Brier score should be reported alongside accuracy for any benchmark with a probabilistic interpretation. ECE has known pathologies (Kumar et al., 2019; Roelofs et al., 2022) — bin count and binning scheme materially change the value, and the standard top-1 estimator misses class-conditional miscalibration — so leaderboards should specify the estimator (equal-mass, equal-width, or kernel-based) or report robustness across binning schemes; Brier and log-score are strictly proper alternatives that avoid the binning question entirely, at the cost of unboundedness on zero-probability predictions for log-score. The structural need is metric-independent: a model correct 70% of the time at uniform 0.95 confidence is over-confident, while a model correct 50% of the time at uniform 0.5 confidence is well-calibrated despite lower accuracy, and only the second is safe to deploy under the deployment asymmetry described under item 1. Accuracy alone cannot distinguish the two. On open-ended generation tasks (Hu-

manEval, GSM8K), the natural probability over outputs is not directly available, and calibration must be measured on derived quantities — semantic-equivalence-cluster probabilities (Kuhn et al., 2023), sampling-based confidence estimates, or LLM-as-judge graded confidence — whose choice is consequential and must be specified per benchmark.

The four changes are pairwise complementary, and we argue informally — at the level a position paper supports rather than as a formal minimality proof — that they are jointly sufficient for the leaderboard-level intervention the proposal aims at. Triple-scoring without abstention-rate reporting cannot be interpreted across models; abstention-rate reporting without triple-scoring is description without incentive; capability-stratified evaluation without triple-scoring has no slice on which to differentially reward refusal; and calibration metrics without abstention reporting cannot distinguish over-confidence from miscalibrated coverage. Whether the bundle is minimal — whether a strict three-item subset would suffice — is a question we leave open; weaker bundles already adopted (HELM, AbstentionBench) have not produced the leaderboard-level gradient the proposal requires.

None of these changes requires new mathematical infrastructure. They require that the scoring function stop rewarding the failure mode the theorem guarantees.

Regulatory framing and fairness audit. The four reforms are not a wishlist; they map onto an existing regulatory framing the field has not yet drawn on. NIST AI RMF Measure 2.9 (National Institute of Standards and Technology, 2023) (validity and reliability) explicitly calls out calibration and uncertainty as measurable system properties, and the EU AI Act's quality-management obligations under Article 15 (European Parliament and Council of the European Union, 2024) are operationalized most directly by exactly the metrics proposed here; a leaderboard that adopted triple-scoring, abstention-rate reporting, and ECE/Brier would produce documentation that meets several of those obligations as a byproduct. The proposal also requires a fairness audit. Triple-scoring incentivizes abstention, and abstention is not symmetric across populations: a model that abstains on Swahili medical QA because it lacks competence is correctly rewarded, but the resulting abstention-rate disparity — between high-resource and low-resource languages (the multilingual leaderboard infrastructure of XNLI, FLORES, and MEGA already provides axes for this), between training-data-represented and underrepresented groups — needs to be reported alongside aggregate abstention rates. Without a stratified abstention-rate audit, triple-scoring converts a population-level information asymmetry into a hidden one.

Transition path and implementation costs. Practically, we expect the lowest-cost path to adoption to mirror HELM's playbook: add triple-scoring and stratified abstention rate as side metrics on existing benchmarks, demonstrate that the resulting rankings differ from accuracy-only rankings in load-bearing ways, and only then promote them to primary. Three implementation costs deserve explicit acknowledgement. First, environment-level unanswerability sets require annotation labor that benchmark maintainers do not currently budget; the model-level case (Item 3) is harder still, and unanswerability annotation has lower inter-annotator agreement than answer-key annotation. Second, leaderboards must adopt a scalarization rule — rank by accuracy plus an abstention bonus, by Brier score, or by a composite weighting — and the choice is consequential and not provided here. Third, scores under triple-scoring are not directly comparable to historical accuracy-only scores, so adoption must include a parallel-reporting transition cycle to preserve longitudinal interpretation across model releases. One asymmetry deserves naming: triple-scoring requires that models emit a structured *abstain* action and that benchmarks recognize it. Proprietary models can be retrained with refusal heads and calibration targets, while open-source releases are typically evaluated post-hoc against a fixed prompt template — so a leaderboard that requires structured abstention will systematically advantage proprietary releases unless the benchmark itself supplies the prompt-template policy that elicits the abstain action.

6 Counterexamples and Boundary Conditions

The strongest objection to the position taken here is that benchmark reform is plainly not a prerequisite for reliability progress, since several deployed model families — including those produced by Anthropic, OpenAI, and Google — have measurably improved abstention behavior through honesty-oriented post-training (Yang et al., 2024) (refusal

training, constitutional methods, calibration heads) without any of the benchmark changes proposed in Section 5. We take the objection seriously. If post-training closes the gap the theorem identifies, the four reforms are optional optimizations rather than structural requirements.

The empirical evidence does not support that reading, but it does support a weaker version of it. Kirichenko et al. (2025) measure a 24% mean degradation in abstention across 20 datasets under reasoning fine-tuning, including in domains the reasoning models were explicitly trained on. Yin et al. (2026)'s reliability-capability trade-off persists across the mitigation strategies they test, with DPO reducing hallucination only by sacrificing utility. Honesty post-training narrows the gap on the slice of inputs where the model itself recognizes uncertainty. We conjecture — and treat this as the testable empirical claim of the position rather than as an established fact — that it does not narrow the gap on the complementary slice the theorem singles out: inputs where consistent reasoning would force confabulation in a region the model believes it knows. The two slices are complementary, the gradients addressing them are different, and the second gradient is the one current benchmarks suppress.

To anchor this empirically, we name a falsification condition: if a post-training intervention without scoring reform achieved sub-10% residual error on a calibrated abstention benchmark stratified by environment-level unanswerability — at the level AbstentionBench currently measures — we would withdraw the structural-necessity reading and treat benchmark reform as a complement rather than a prerequisite for principled closure of the theorem-identified class.

We therefore do not claim that post-training is mistaken or that the field should abandon it. The narrower position is that benchmark reform is necessary, not sufficient, for closing the class Bastounis identifies, and post-training is necessary, not sufficient, for closing the class users most often encounter. The two interventions address different gradients, and a reliability program that pursues only one will plateau at the boundary of the other.

Limitations

We do not claim the convergence implies that the three papers' specific interventions or proposed mitigations are equivalent. Yin's work points at activation-level diagnostics. Suleymanov's points at agentic scaffolding for refinement. Bastounis's points at the "I don't know" function as a recursion-theoretic object. These remain distinct research programs even if the warranted intervention is the same.

We do not claim our four-part evaluation proposal is the only possible response to the diagnosis. A reasonable alternative would propose mechanism-level interventions (calibration heads, refusal tokens, conformal layers) that retrofit existing benchmarks rather than reforming them. We argue that without scoring reform such interventions are selected against rather than for.

We do not claim our convergence reading is the only possible reading of Yin, Suleymanov, and Bastounis. Other accounts may fit the three papers equally well. We claim that treating them as three separate reliability problems misses what they share, and that what they share has policy consequences the field has not yet acted on.

We do not claim the three findings represent the same kind of unavailability. Yin's is a capability gap; Suleymanov's a policy gap; Bastounis's theorem is agnostic to source — which is what lets the three findings converge on the same remedy despite having different origins. The distinction matters for evaluation design: capability gaps can be probed with tasks that are ground-truth unanswerable, policy gaps require benchmarks that specify the ruling-out rule. Our Section 5 proposal addresses the first cleanly and gestures at the second.

We do not claim the proposal generalizes to long-form generation. For summaries, multi-paragraph answers, and multi-step reports, calibrated abstention has no clean discrete analogue at the document level, and operationalizing triple-scoring there is an open problem outside this proposal's scope.

References

Alexander Bastounis, Paolo Campodonico, Mihaela van der Schaar, Ben Adcock, and Anders C. Hansen. 2024. On the consistent reasoning paradox of intelligence and optimal trust in ai: The power of 'i don't know'. *Preprint*, arXiv:2408.02357.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

C. K. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. Commonly known as the EU AI Act; Article 15 covers accuracy, robustness, and cybersecurity.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*. ArXiv:2009.03300.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696. Association for Computational Linguistics.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems 38: Datasets and Benchmarks Track*. ArXiv:2506.09038.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Ananya Kumar, Percy S. Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*. ArXiv:2211.09110, originally posted Nov 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2305.14251.

National Institute of Standards and Technology. 2023. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Sello Ralethe and Jan Buys. 2025. KERLQA: Knowledge-enhanced reinforcement learning for question answering in low-resource languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1834–1846, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. 2022. Mitigating bias in calibration error estimation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Leonard J. Savage. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Umid Suleymanov, Zaur Rajabov, Emil Mirzazada, and Murat Kantarcioglu. 2026. Beyond refusal: Probing the limits of agentic self-correction for semantic sensitive information. *Preprint*, arXiv:2602.21496.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen tau Yih, and Xin Luna Dong. 2025. TruthRL: Incentivizing truthful LLMs via reinforcement learning. *Preprint*, arXiv:2509.25760.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556. ArXiv:2407.18418.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS '24)*, Red Hook, NY, USA. Curran Associates Inc.

Chenlong Yin, Zeyang Sha, Shiwen Cui, Changhua Meng, and Zechao Li. 2026. The reasoning trap: How enhancing LLM reasoning amplifies tool hallucination. In *International Conference on Learning Representations (ICLR)*. ArXiv:2510.22977.

Skylar Zhai, Jingcheng Liang, and Dongyeop Kang. 2026. Abstain-R1: Calibrated abstention and post-refusal clarification via verifiable RL. *Preprint*, arXiv:2604.17073.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.