Position: When Do Benchmark Scores Predict Deployment? A Validity Framework for Capability Claims

Dipam Paul [1]

Abstract

Frontier models achieving 75.7% on ARC-AGI-1 collapse to 3–4% on structurally novel ARC-AGI-2—a 95% relative score degradation—yet rankings remain stable ($\tau > 0.9$). This pattern recurs: SWE-bench scores drop from >70% to 23% on novel repositories (rankings preserved); score jumps on MMLU and GSM8K (+7–17 pts) correlate with training-cutoff shifts without corresponding deployment gains—yet diagnosing these failures is difficult because training data composition is proprietary and contamination detection fails for reasoning models. We argue that these failures arise from conflating two epistemically distinct claim types: *method-ranking claims* (reliable without validity evidence) and *capability claims* (requiring scope evidence before deployment). Drawing on measurement theory, we formalize task structure preservation criteria predicting when rankings transfer versus when scores collapse. For practitioners, we propose the Robustness Ratio ($RR = \sigma^2_{\text{paraphrase}}/\sigma^2_{\text{seed}}$) as a falsifiable diagnostic: two frontier models both scoring 100% on 50 MMLU questions reveal a $\sim 40\times$ gap in paraphrase sensitivity—a difference invisible to headline scores. We also propose tiered validity protocols adding only 1.2–2× evaluation cost. Our framework specifies what would refute it: if matched-score models show equivalent deployment performance regardless of training data composition, our distinction is wrong.

1. Introduction

The evaluation timeline is compressing. Benchmarks designed to last years saturate in months. FrontierMath, released November 2024 with PhD-level problems estimated to "take years," improved from 2% to 10–25% within months (Epoch AI, 2025). ARC-AGI-2, designed to resist memorization, saw Poetiq achieve 75% on the public set by December 2025 (ARC Prize Foundation, 2025a). The Common Task Framework has driven reproducible progress for decades (Donoho, 2017), but models now exceed human baselines on reasoning benchmarks yet fail consistently in deployment (Raji et al., 2021).

Our position. Benchmark performance should be interpreted as evidence about *methods*—algorithmic innovations and training procedures—rather than direct evidence of model *capabilities*. Within-cohort rankings remain stable for comparing algorithms (Salaudeen & Hardt, 2024); capability claims require additional validity evidence.

Deployment failures. Coding agents achieving high SWE-bench scores produce erroneous patches and security vulnerabilities in deployment (Tao et al., 2024). Medical AI with >90% benchmark accuracy shows 20–35% degradation on underrepresented populations (Raji et al., 2021). These failures share a pattern: benchmark success predicts deployment on benchmark-similar inputs but not on real-world distributions. Understanding *why* requires distinguishing problem types.

The core distinction. Benchmark-deployment gaps arise from multiple causes. We distinguish three problem types:

Problem	Type	Relevance
Validity	Construct	**Core**
Measurement	Quality	Fixable
Precision	Quality	Fixable

Measurement quality and precision are engineering problems with known solutions (Tao et al., 2024; Yao et al., 2024). Our focus is *construct validity*—whether benchmarks measure what they claim to measure. Benchmarks with perfect measurement fail to predict deployment when they measure performance on training-similar tasks (interpolation) rather than genuinely novel tasks (extrapolation).

We distinguish two failure modes: (1) *Construct invalidity*—the benchmark does not measure its claimed property; (2) *Scope violation*—the benchmark validly measures a property but inferences extend beyond warranted scope. The

[1]ServiceNow Inc.. Correspondence to: Dipam Paul <dipam.paul@servicenow.com>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

SAT illustrates this: scores predict college GPA ($r \approx 0.48$–0.51; Westrick et al. 2019) but not physician performance—a scope violation, not invalidity. Most evidence in Section 3 demonstrates scope violations.

Scope. Our critique applies to *capability claims*—that models "reason" or "generalize" beyond benchmark distributions. Method-ranking claims require reliability; capability claims require scope evidence. "GPT-4 scores 86% on MMLU" is a benchmark report; it becomes a capability claim when used to infer "GPT-4 reasons reliably on novel tasks." This aligns with Messick's validity theory: validity concerns the inferences drawn from scores, not the test itself (Messick, 1989; Kane, 2006).

Contributions. We make four contributions:

- **Conceptual framework:** We formalize the distinction between *method-ranking claims* (requiring reliability evidence) and *capability claims* (requiring validity evidence), grounding this in measurement theory (Section 2).
- **Empirical synthesis:** We aggregate evidence showing rankings transfer when task structure is preserved (Kendall's $\tau > 0.9$), but scores are confounded by training overlap (+7–17 pts) and collapse under structural novelty (ARC-AGI: 75%→4%; SWE-bench Pro: 70%→23%) (Section 3).
- **Actionable protocols:** We propose the Robustness Ratio (RR) and tiered validity protocols (1.2–2× cost) with specific thresholds for practitioners (Section 5).
- **Falsifiable predictions:** We specify testable conditions under which our framework would be refuted, including $\tau > 0.8$ for structure-preserving variants vs. $\tau < 0.5$ for structure-altering variants (Section 2).

The current state. Evidence indicates models optimize for evaluations beyond contamination, though direct causal evidence remains limited. LLMs guess masked MMLU options with 52–57% exact match accuracy (Deng et al., 2024); GameArena documents benchmark-specific behaviors (Hu et al., 2025). By January 2026, frontier models achieve near-ceiling performance on MMLU (Hendrycks et al., 2021) and similar benchmarks (Table 1).

Saturation admits two interpretations: genuine extrapolation capability or shared interpolation ceiling. The distinction is testable: the ARC-AGI collapse (o3's 75.7% on ARC-AGI-1 vs. ∼4% on ARC-AGI-2) and SWE-bench Pro degradation (>70% → 23%) are consistent with interpolation dominance, though alternative explanations cannot be fully ruled out (Scale AI, 2025).

Epistemic status. Our confidence varies by claim type:

- *High confidence*: Vision ranking transfer (Kendall's $\tau >$ 0.9 to novel distributions), contamination effects (+7–17 points per Domínguez-Olmedo et al. 2025)

Table 1. Benchmark saturation as of January 2026. Frontier model performance on major benchmarks. The ARC-AGI-1→ARC-AGI-2 collapse (75.7%→3–4%) motivates our framework.

Benchmark	'24	'26	Notes
MMLU	86%	92–95%	Frontier
GSM8K	92%	97–99%	Reasoning
FrontierMath	2%	10–25%*	2-mo sat.
ARC-AGI-1	5%	75.7%†	o3-high
ARC-AGI-2	—	3–4%‡	o3 baseline

*OpenAI: 25%; Epoch: 10%. †Original benchmark. ‡Harder version; Poetiq: 54–75%.

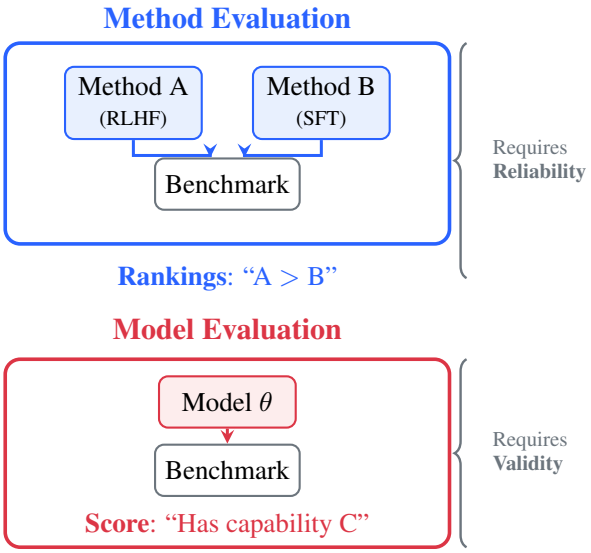

Figure 1. **The core distinction.** Method evaluation (top) compares training procedures (e.g., RLHF vs. SFT) via *rankings* ("A > B"), requiring only reliability (consistency). Model evaluation (bottom) assesses whether a model possesses a capability via *absolute scores*, requiring validity (evidence that scores predict the claimed property). Rankings survive noise; capability claims require transfer evidence.

- *Moderate confidence*: LLM ranking transfer under paraphrase ($\tau > 0.9$), though vision-to-LLM extrapolation requires caution (Section 3)
- *Uncertain*: Test-time compute (whether search constitutes extrapolation remains open)
- *Limited*: Reasoning model contamination detection (emerging methods show promise)

Our critique applies strongly to ill-defined constructs ("reasoning"), weakly to verifiable benchmarks (code execution), and minimally to saturated benchmarks. See Appendix G for detailed confidence levels.

Figure 1 illustrates our core distinction. Section 2 formalizes the framework, Section 3 presents evidence, Section 4 addresses objections, and Section 5 proposes concrete steps.

2. The Core Distinction

We formalize the distinction between method and model evaluation, drawing on measurement theory (Jacobs & Wallach, 2021; Nunnally & Bernstein, 1994) and recent ML evaluation critiques (Raji et al., 2021; Liao et al., 2021).

2.1. Definitions

Definition 2.1 (Method)**.** A *method* is the procedure for creating a model: architecture, optimization algorithm, data curation strategy, and hyperparameters. Our distinction is *epistemological*: method claims require reliability evidence, capability claims require validity evidence (see Appendix H).

Definition 2.2 (Model)**.** A *model* is the weights $\theta \in \mathbb{R}^d$ resulting from executing a method. Deployment claims for systems with scaffolding (RAG, agents) must assess the complete system.

Definition 2.3 (Capability)**.** A model possesses a *capability* C **to degree** d if it reliably exhibits behavior B across a distribution $\mathcal{D}$ relevant to C, where d is determined by: (1) **Generalization**—$\mathcal{D}$ includes inputs outside training distribution; (2) **Robustness**—stable under paraphrase and format changes; (3) **Consistency**—reproducible across trials. **Normative status:** This definition is a *proposed criterion* for when capability claims are warranted, not an analytic truth—we argue these criteria should be required, a position defended by evidence in Section 3. For deployment contexts, additional criteria may include alignment (behavior matches intent) and calibration (confidence matches accuracy). Capability is a continuum: practical thresholds depend on deployment stakes.

Definition 2.4 (Interpolation vs. Extrapolation)**.** Performance decomposes into two components:

- *Interpolation*: Performance on tasks within the support of the training distribution—same task structure, similar surface statistics.
- *Extrapolation*: Performance on tasks outside this manifold—novel task structure, different statistical patterns.

Operational test. Given benchmark B, create variant B' that changes task format or domain mapping while preserving the underlying construct. Interpolation-dominant models degrade sharply on B'; extrapolation-capable models maintain performance. Evidence: MMLU $\rightarrow$ MMLU-Pro (preserved structure, $\tau > 0.9$; Wang et al. 2024) vs. ARC-AGI-1 $\rightarrow$ ARC-AGI-2 (altered structure, 75.7% $\rightarrow$ 3–4%; ARC Prize Foundation 2025a).

Disentangling difficulty from structural novelty. The WILDS benchmark (Koh et al., 2021) provides supporting evidence: mixed-split training (combining in-distribution and out-of-distribution data) substantially reduces performance gaps, suggesting that degradation stems primarily from distribution shift rather than intrinsic difficulty. The ARC-AGI-2 collapse provides complementary evidence: humans maintain 60–75% accuracy while AI drops from 75.7% to 3–4% (ARC Prize Foundation, 2025a)—if the collapse were difficulty-driven, human performance should degrade proportionally (Chollet, 2019). Compositional-ARC (Hsu et al., 2025) shows a 5.7M-parameter model trained via meta-learning significantly outperforms o3-mini on compositional generalization, providing evidence that structural novelty (not scale or difficulty) determines transfer.

Avoiding tautology. The distinction is not about generalization per se—both interpolation and extrapolation generalize. The distinction concerns *what* generalizes: interpolation transfers across distributions with shared statistical structure; extrapolation transfers across distributions with different structure but shared semantic content.

Definition 2.5 (Scope vs. Validity Failure)**.** Building on the distinction introduced in Section 1: *construct validity failure* means scores do not correlate with the claimed construct even within-distribution; *scope violation* means scores predict well within-distribution but fail out-of-distribution. Construct validity failure suggests redesign; scope violation suggests respecting bounds. Most evidence in Section 3 demonstrates scope violations (Cronbach & Meehl, 1955).

Definition 2.6 (Task Structure)**.** Task structure comprises: (1) input format, (2) reasoning depth, (3) domain vocabulary overlap, (4) solution template. A variant B' *preserves structure* if these criteria are met:

Operationalization: Format unchanged; reasoning depth ratio $\in [0.5, 2.0]$; vocabulary overlap PAD < 0.5 or Jaccard > 0.7 (Ben-David et al., 2006); template unchanged. See Appendix L for threshold derivation.

Validation metric (not definitional): We use ranking correlation τ as an *empirical test* of structure preservation, not as part of the definition. The criteria above specify structure preservation a priori; τ provides post hoc validation. **Prediction**: structure-preserving variants (per the above criteria) should show $\tau > 0.8$; structure-altering variants should show $\tau < 0.5$. This prediction is falsifiable. See Appendix L for threshold derivation and Appendix M for diagnostics.

2.2. Variance Decomposition Framework

We propose an operational framework for interpreting benchmark scores. For any benchmark B and model population M, total score variance can be partitioned:

$$\text{Var}(\text{Score}) = \text{Var}_{\text{interp}} + \text{Var}_{\text{extrap}} + \text{Var}_{\text{game}} + \text{Var}_{\text{noise}}$$

where:

- $\text{Var}_{\text{interp}}$ ($\hat{\alpha}$): variance explained by training-test distribution overlap
- $\text{Var}_{\text{extrap}}$ ($\hat{\beta}$): variance explained by transfer to structurally novel tasks
- Var_{game} ($\hat{\gamma}$): variance from models actively optimizing for evaluation patterns
- $\text{Var}_{\text{noise}}$: measurement noise, prompt sensitivity, random seeds

Estimation. Direct estimation requires contamination metrics unavailable for closed models. Proxies: (1) paraphrase degradation for $\hat{\alpha} + \hat{\gamma}$; (2) temporal discontinuity for $\hat{\alpha}$; (3) structure-transfer for $\hat{\beta}$. The decomposition is operational rather than ontological (Appendix E).

Predictions. (P1) Static benchmarks have high $\hat{\alpha}+\hat{\gamma}$, low $\hat{\beta}$. (P2) Interactive benchmarks have higher $\hat{\beta}$. (P3) Evaluation-aware models have high $\hat{\gamma}$. We include Var_{game} because peer-reviewed work documents active evaluation optimization (Deng et al., 2024; Hu et al., 2025).

Quantitative evidence. Recent work provides empirical grounding: GSM8K shows up to 13% accuracy inflation attributable to contamination (Zhang et al., 2024); Llama 2-70B exhibits 18.1% n-gram overlap with benchmark data (Shi et al., 2025). For $\hat{\gamma}$, evaluation-awareness detection achieves AUC = 0.83 for frontier models, with detection ability scaling as a power law with model size (Needham et al., 2025). The DCR framework (Xu et al., 2025) provides contamination-adjusted accuracy within 4% average error, demonstrating practical estimability.

2.3. Reliability vs. Validity

From measurement theory (Nunnally & Bernstein, 1994; Cronbach & Meehl, 1955) (see Appendix D for detailed comparison with ML evaluation):

Reliability is consistency under perturbation—same relative ordering across measurements. **Validity** is whether measurement captures the intended property. Reliability is necessary but not sufficient for validity; a perfectly consistent measurement can consistently measure the wrong thing.

Relationship to evaluation science. Our framework operationalizes triangulated measurement (Weidinger et al., 2025): within-cohort rankings provide internal validity; transfer tests provide external validity; variance decomposition addresses construct validity.

Identifiability. Interpolation and extrapolation may not be cleanly separable, but the epistemological distinction—method claims require reliability, capability claims require validity—holds regardless (Appendix E).

Table 2. Method vs. model evaluation: key differences.

Aspect	Method Eval.	Model Eval.
Target	Algorithm	Weights θ
Output	Rankings	Absolute scores
Requirement	Reliability	Validity
Noise-robust	Yes	No

2.4. Why Rankings Survive Noise

Rankings survive strictly increasing transformations (Proposition A.1); symmetric noise contracts scores toward a constant while preserving order (Proposition A.2; proofs in Appendix A). **Critical limitation:** These apply only to *symmetric* perturbations. Contamination is *asymmetric*—MMLU-CF (Zhao et al., 2025b) shows rankings "changed considerably" on contamination-free reformulations (see Appendix F for mechanism analysis). Within-cohort rankings survive noise; cross-cohort comparisons are confounded.

What would falsify this framework? Our position would be refuted if: (1) benchmark scores predicted deployment performance equally regardless of distribution similarity; (2) fine-tuning equalization failed to reduce score gaps; (3) matched-score models showed equivalent held-out performance regardless of training data composition. Evidence supports our framework (Section 3).

Three regimes, not two. Our framework is sometimes read as a binary (rankings survive, scores collapse). The empirical picture has three regimes: *structure-preserving* variants show stable rankings and scores (MMLU paraphrase: $\tau > 0.9$, <5% drop); *partial-shift* variants show score drops with preserved rankings (MMLU→MMLU-Pro: −16–33% drop, $\tau > 0.9$); and *structure-altering* variants show score collapse with rankings unmeasurable at floor (ARC-AGI-1→2: 75.7%→3–4%). "Collapse" in our title and abstract refers to score collapse in the third regime, not ranking collapse. This distinction is critical: RR is most diagnostic in the partial-shift regime, where rankings survive but paraphrase sensitivity reveals hidden fragility that leaderboards destroy.

3. Empirical Evidence

We test three hypotheses derived from our framework:

- **H1 (Reliability)**: Rankings remain stable ($\tau > 0.8$) under structure-preserving transformations.
- **H2 (Interpolation)**: Absolute scores inflate with training-test overlap.
- **H3 (Extrapolation)**: Performance collapses (>50% relative degradation) under structural novelty.

We organize evidence by hypothesis; Table 3 summarizes

Table 3. Empirical evidence summary. Type codes: *Reliab.* = reliability (H1); *Interp.* = interpolation (H2); *Extrap.* = extrapolation failure (H3). *MMLU-CF rankings "changed considerably" across cohorts (Zhao et al., 2025b).

Finding	Effect	Type
Paraphrase ranking*	τ>0.9	Reliab.
MMLU→Pro	−16–33%	Reliab.
MMLU inflation	+7 pts	Interp.
GSM8K inflation	+17 pts	Interp.
ARC-AGI-1→2	75%→4%	Extrap.
SWE-bench Pro	70%→23%	Extrap.

key findings, with detailed analysis in each subsection below.

3.1. H1: Ranking Stability Under Structure-Preserving Transformations

Vision vs. LLM evidence: a critical distinction. Our evidence base differs qualitatively between modalities, and we acknowledge this asymmetry explicitly:

Vision evidence (strong). ImageNot (Salaudeen & Hardt, 2024) reports $\tau > 0.9$ ranking transfer to an *entirely different dataset*—same task (image classification), novel images with different statistical properties. Recht et al. (2019) report strong ranking preservation between ImageNet and ImageNet-V2 despite 11–14% absolute accuracy drops. Vision rankings transfer across distributions with preserved task structure.

LLM evidence (moderate → strong under paraphrase). Lunardi et al. (Lunardi et al., 2025) tested 34 LLMs across 264,761 paraphrased questions, finding $\tau > 0.9$ **for all six benchmarks** (ARC-C, HellaSwag, MMLU, OpenBookQA, RACE, SciQ; all $p < 0.01$). Benchmark2 (Qian et al., 2025) reports cross-benchmark ranking consistency with average $\tau = 0.93$ using selective evaluation. Mathematics benchmarks show particularly strong consistency ($\tau = 0.88$–0.99 between OlympiadBench, OmniMath, MATH-500, AMC). This suggests comparable ranking stability, though the nature of the shift differs: vision evidence involves genuinely novel images, while LLM evidence involves linguistic reformulations of the same questions. MMLU → MMLU-Pro shows $\tau > 0.9$ despite 16–33% score drops (Wang et al., 2024); LiveBench reports $\tau > 0.997$ rank correlation between consecutive monthly updates (White et al., 2024). However, rankings are stable under *surface-level* variations (paraphrase, format) but sensitive to *methodological* changes.

Counter-evidence and boundary conditions. Thomas et al. (2024) document up to 8 position shifts from prompt template changes alone. MMLU-CF (Zhao et al., 2025b) shows rankings "changed considerably" on contamination-free reformulations (GPT-4o: 88%→73%)—demonstrating that cross-cohort comparisons are confounded by asymmetric data exposure.

Why vision-to-LLM extrapolation requires caution. We explicitly acknowledge that extrapolating from vision ranking transfer to LLMs is not straightforward:

- *Different failure modes*: Vision models fail primarily due to distribution shift; LLMs fail due to contamination, prompt sensitivity, and direct benchmark optimization during training.
- *Different test conditions*: Vision evidence (ImageNot, ImageNet-V2) involves genuinely novel images never seen during training; LLM paraphrase evidence involves linguistic reformulations of potentially contaminated questions.
- *Different saturation dynamics*: ImageNet rankings remain discriminative; LLM benchmarks show ceiling effects (frontier models clustered above 88% on MMLU) reducing discriminative power.

Our framework's predictions for LLMs are therefore *hypotheses requiring validation*, not established facts extrapolated from vision. The GSM8K-Novel experiment—math problems with valid syntax but genuinely novel structure—would test whether LLM rankings transfer as robustly as vision rankings (estimated: ∼50× baseline cost).

Implication. Rankings are reliable for method comparison under structure-preserving variations, with stronger evidence for vision than LLMs. For LLMs, "structure-preserving" currently means linguistic paraphrase with preserved semantics—a weaker test than vision's transfer to novel distributions.

3.2. H2: Score Inflation from Training-Test Overlap

Domínguez-Olmedo et al. (2025) found models trained after November 2023 showed jumps on MMLU (+7 pts) and GSM8K (+17 pts) without scale increases—evidence consistent with high $\hat{\alpha}$. Within-cohort rankings remain stable; cross-cohort rankings are confounded. **Clarification.** Contamination is not a competing construct—it is the primary empirical mechanism by which interpolation dominance manifests. The interpolation framing explains *why* contamination inflates scores (models exploit training-similar surface patterns) while leaving within-cohort rankings intact (relative ordering is preserved under near-uniform inflation). This unifies H2 (score inflation) and H3 (structural collapse) within the same framework. This inflation raises the question of whether contamination can be reliably detected—current evidence suggests not.

Contamination detection is failing. For reasoning models (o1, o3, DeepSeek-R1), RL training (GRPO/PPO) conceals contamination signals that probability-based methods rely

Table 4. SWE-bench performance across contamination-resistant test sets (Scale AI, 2025). Degradation demonstrates interpolation effects persist even with execution verification.

Model	Verified	Pro (Pub.)	Pro (Priv.)
Frontier Model A	>70%	23.3%	14.9%
Frontier Model B	>70%	22.7%	17.8%
GPT-4o	—	4.9%	—

on, with detection accuracy estimated to fall near chance level (see Appendix B for methodology). This limitation: as reasoning models become the frontier, traditional validation methods fail. Emerging approaches show promise—Self-Critique methods detecting entropy collapse achieve up to 30% AUC improvement; output distribution analysis (CDD) provides 21–30% relative improvement over baselines—but these require validation on reasoning models specifically. For now, we recommend: (1) dynamic benchmarks (LiveBench, LiveCodeBench) that avoid contamination by construction, and (2) behavioral validation via perturbation testing rather than membership inference. See Appendix B for detection method comparison.

3.3. H3: Extrapolation Failure Despite Verification

For benchmarks where correctness is verified via code execution (e.g., test suite execution), our critique applies less strongly; these benchmarks provide ground truth that multiple-choice formats lack. However, SWE-bench Pro (Scale AI, 2025) provides strong evidence that interpolation effects persist even with verifiable outputs (Table 4).

The degradation is consistent with bounded predictive scope: SWE-bench Pro uses GPL/copyleft repositories that may be underrepresented in commercial pre-training corpora, though this hypothesis requires verification. This correlation supports our interpolation hypothesis, though we cannot rule out alternative explanations such as differing code complexity or repository characteristics. Execution verification strengthens but does not eliminate interpolation/extrapolation confounds.

Summary. The evidence is consistent with all three hypotheses: H1 (ranking stability, $\tau > 0.9$ under structure-preserving transformations), H2 (score inflation, +7–17 pts from training overlap), and H3 (extrapolation failure, ARC-AGI: 75%→4%; SWE-bench Pro: 70%→23%). Counter-evidence exists for H1 (MMLU-CF ranking changes, prompt sensitivity) and H3 evidence comes from limited benchmarks; further validation is warranted.

The evidence supports our core distinction: rankings reliably compare methods (H1), but absolute scores are confounded by training overlap (H2) and collapse under structural novelty (H3). This framing raises natural objections, which we address next.

4. Alternative Views and Responses

4.1. "Don't Benchmarks Already Predict Deployment?"

Objection. Miller et al. (2021) showed "accuracy on the line"—benchmark predicts deployment.

Response. This finding holds under mild distribution shift. Teney et al. (2023) clarify the mechanism: positive in-distribution/out-of-distribution correlations occur when task structure is preserved (the OOD data shares statistical regularities with training data), but inverse correlations emerge when structure is disrupted. Benchmarks predict well when deployment matches benchmark distribution, task structure is preserved, and models are paraphrase-robust. We claim their scope is bounded, not that benchmarks are invalid (Figure 2).

4.2. "Interpolation IS Capability at Scale"

Objection. For narrow commercial deployments, interpolation performance IS the product. At sufficient scale, the interpolation manifold may cover nearly all scenarios.

Response. We agree for benchmark-similar deployments. However, SWE-bench Pro's 47-point degradation on legally-excluded repositories suggests interpolation gaps persist at frontier scale. If ARC-AGI-2 is dismissed as adversarial, doing so implicitly acknowledges that deployment scenarios outside the interpolation manifold exist. The question is frequency—requiring deployment studies. **Prediction**: if this objection holds, frontier models should show modest degradation (not collapse) on held-out deployments. Current evidence (SWE-bench Pro: >70%→23%, a 67% relative drop; ARC-AGI: 75%→4%, a 95% relative drop) shows collapse rather than modest degradation, supporting our framework.

4.3. "Doesn't Test-Time Compute Enable Genuine Reasoning?"

Objection. Reasoning models achieve gains through inference-time scaling. Isn't search a genuine capability?

Response. Test-time compute is a method improvement (detailed analysis in Appendix I). The ARC-AGI collapse and ARC-AGI-3 results (best agent: 12.6% vs. human performance approaching ceiling; ARC Prize Foundation 2025b) suggest reasoning strategies don't transfer to structurally novel tasks. We hypothesize that reasoning in current models is interpolation-dependent: models search solution spaces resembling training structure. We acknowledge that task difficulty and structural novelty are partially confounded, and increased difficulty is a plausible

alternative explanation. However, two observations favor the structural-novelty interpretation: (1) the magnitude of collapse (75.7%→3–4%) exceeds typical difficulty-driven degradation, and (2) humans maintain 60–75% accuracy on ARC-AGI-2 (ARC Prize Foundation, 2025a), suggesting the tasks are not intrinsically unsolvable.

4.4. "The Framework Is Circular"

Objection. The framework appears tautological: you define extrapolation as transfer to novel structure, then use transfer failure as evidence for interpolation. How is this falsifiable?

Response. The circularity concern is valid if task structure is defined *post hoc* by transfer results. We avoid this via *a priori* criteria (Definition 2.6): format preservation, reasoning depth ratio, vocabulary overlap, and template preservation are specified *before* observing transfer. The framework makes falsifiable predictions: structure-preserving variants (per these criteria) should show $\tau > 0.8$; structure-altering variants should show $\tau < 0.5$. If a benchmark pair satisfies all structure-preservation criteria yet shows $\tau < 0.5$, or violates criteria yet shows $\tau > 0.8$, our framework is refuted. Current evidence (MMLU→MMLU-Pro: structure-preserving, $\tau > 0.9$; ARC-AGI-1→2: structure-altering, 95% collapse) is consistent with predictions, but the framework invites adversarial testing.

4.5. "Are Rankings and Capability Claims Fully Separable?"

Objection. Ranking claims carry implicit capability connotations: stating "Method A outperforms Method B on coding benchmark X" arguably implies both methods have some coding capability.

Response. We agree that rankings often carry implicit capability connotations, and we do not claim full practical separability. Our distinction is *epistemological*: the *evidence requirements* differ. A ranking claim requires only reliable ordering—which contamination preserves within-cohort (H1). A capability claim requires evidence that scores transfer beyond the benchmark distribution (H3). In practice: a ranking tells a researcher which training procedure to prefer; a capability claim tells a practitioner whether to deploy. These are distinct decisions with different risk profiles and different evidence standards. When two methods score identically, RR distinguishes the one with robust paraphrase generalization—a decision ranking alone cannot support, and one where the implicit capability connotation matters.

From diagnosis to action. These objections clarify rather than undermine our framework: benchmarks predict deployment when distributions match (first objection), interpolation suffices for benchmark-similar deployments (second objection), but neither test-time compute (third objection) nor post-hoc structure definitions (fourth objection) resolve the interpolation-extrapolation confound. The question is not whether benchmarks are useful, but how practitioners can determine when capability claims are warranted. Section 5 provides concrete protocols.

5. From Diagnosis to Action

5.1. Evidence Requirements by Claim Type

Capability claims are scope declarations—implicit warranties for what models can do. Method claims require reliability evidence; interpolation claims require $RR < 1.5$ (preliminary threshold; see Section 6); extrapolation claims require transfer tests on $\geq$3 held-out distributions. A *held-out distribution* is operationally defined as: (1) data collected after model training cutoff, or (2) data from domains with $<$10% vocabulary overlap with known training sources, or (3) data violating structure-preservation criteria per Definition 2.6. We acknowledge that constructing genuinely held-out distributions is resource-intensive; for resource-constrained settings, existing OOD benchmarks (e.g., MMLU-Pro, SWE-bench Pro) may serve as proxies. Validity requirements should scale with capability-deployment mismatch (Appendix C). Given these evidence requirements, we now provide actionable guidance for different stakeholders.

5.2. Recommendations

For researchers. Append claim type: *[Method]*, *[Interpolation]*, *[Extrapolation-Preliminary/Validated]*. **For venues.** Pilot a "Validity-Checked" track with Table 5 tiers. **For practitioners.** Request Robustness Ratio before procurement where available; for extrapolation deployments, request transfer test results from vendors or conduct independent pilot evaluations on deployment-representative samples (Appendix O).

Bridging the audience gap. Procurement teams, policy makers, and deployment engineers—the audiences most reliant on benchmark scores—seek capability claims but often lack resources for full transfer-test protocols. For these audiences, the minimum viable step is: *request RR alongside any benchmark score before high-stakes deployment*. Vendors capable of reporting benchmark scores can compute RR with marginal additional overhead ($\sim 6\times$ the base API calls; see Appendix N). This single number distinguishes two models with identical headline scores but $40\times$ different paraphrase fragility—a difference that matters for deployment but is invisible to leaderboards. Importantly, ranking claims carry implicit capability connotations ("Method A outperforms B on coding benchmark" implies some coding competence exists); RR addresses precisely those cases

where identical rankings hide capability differences relevant to deployment.

Domain-specific guidance. For *code benchmarks* (SWE-bench, HumanEval): execution verification provides stronger validity, but SWE-bench Pro's 47-point degradation on legally-excluded repositories suggests interpolation effects persist. For *math benchmarks* (GSM8K, MATH): contamination-adjusted evaluation (GSM1k, MATH-Perturb) shows up to 16% performance drops on hard perturbations (Zhang et al., 2024; Fan et al., 2025). Rankings remain stable ($\tau > 0.88$) within math benchmarks (Qian et al., 2025), but capability claims require transfer to novel problem formats.

Scope caveat. For benchmark-similar deployments, interpolation benchmarks serve their commercial purpose. Our critique targets capability claims in marketing, deployment to novel distributions, and safety-critical contexts.

6. Toward Capability-Measuring Benchmarks

Having proposed evidence requirements, we now turn to constructive proposals for improving benchmark validity. Benchmarks like SWE-bench-Live and LiveCodeBench (Jain et al., 2024) implement continuous updates. We propose extending these with: (1) temporal freshness (using only post-training-cutoff data) with verification; (2) adversarial resistance via paraphrased variants; (3) multi-trial evaluation; (4) cost-adjusted metrics; (5) compute-stratified evaluation (reporting performance at multiple inference budgets) for reasoning models. ARC-AGI-3 (ARC Prize Foundation, 2025b) pioneers interactive evaluation (best agent: 12.6% vs. near-perfect human), increasing $\hat{\beta}$ (extrapolation variance) by requiring capabilities that cannot be memorized.

The Robustness Ratio. We propose $RR = \sigma^2_{\text{paraphrase}}/\sigma^2_{\text{seed}}$, where σ^2_{seed} is variance across random seeds on identical prompts and $\sigma^2_{\text{paraphrase}}$ is variance across semantically equivalent reformulations. With *preliminary* thresholds derived from observed variance ratios in Lunardi et al. (2025): $RR < 1.5$ suggests robust performance (paraphrase variance within 50% of seed variance); $RR > 3.0$ suggests fragile pattern-matching (paraphrase variance dominates). These thresholds are working hypotheses requiring empirical validation before deployment use; see Appendix N for the validation protocol. **Important caveat**: RR is a measurable *correlate* of fragile pattern-matching, not direct proof of training-test overlap—the latter cannot be externally verified for closed-source frontier models. High RR indicates a model whose scores are paraphrase-sensitive; the mechanism (contamination, distribution mismatch, or other factors) remains an open question. An illustrative case: Claude Sonnet 4 and Gemini 2.5 Flash both scored 100% on 50 MMLU originals yet differed $\sim 40\times$ in $\sigma^2_{\text{paraphrase}}$; this was the complete model set evaluated, not a post-hoc selection (see Appendix N for model selection rationale).

Table 5. Tiered validity protocols. Cost = API/compute multiplier; paraphrase generation requires ∼40h one-time researcher effort (author estimate).

Claim Type	Cost	Requirements
Method ranking	1.0×	Standard benchmark
Interpolation	1.2×	+ paraphrase (RR<1.5)
Extrap. (prelim.)	1.5×	+ multi-format eval
Extrap. (valid.)	2.0×	+ transfer (≥3 dists.)

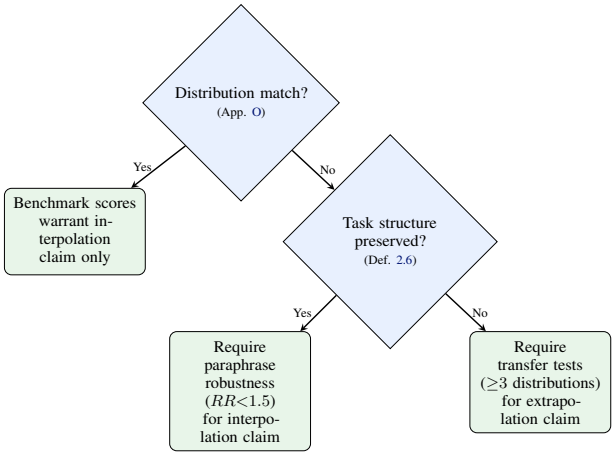

Figure 2. **Decision tree for practitioners.** Evidence requirements scale with deployment-benchmark mismatch (Appendix O). *Example*: SWE-bench-evaluated assistants deployed to enterprise codebases with different licensing require transfer tests.

Tiered validity protocols. Beyond the RR diagnostic, Table 5 presents cost-stratified requirements accessible to academic labs.

Figure 2 operationalizes these protocols as a practitioner decision tree.

Goodhart caveat. These validity protocols will themselves become optimization targets once widely adopted. A specific concern: if paraphrasing becomes standard evaluation, models may be trained on paraphrased data (data augmentation), treating variance rather than understanding. We address this via dynamic paraphrase generation and defense-in-depth (Appendix N). We view evaluation as an adversarial game requiring continuous innovation rather than a solvable problem; our proposals buy time but are not permanent solutions. See Appendix J for extended discussion.

7. Related Work

Construct validity critiques. Raji et al. (2021) documented benchmark-deployment gaps; Sherborne et al. (2025) taxonomized validity failures; Alaa et al. (2025) showed benchmark scores fail to predict clinical outcomes. We formalize these via the method-ranking vs. capability-claim distinction.

Contamination. Domínguez-Olmedo et al. (2025) quantified contamination effects; Zhao et al. (2025b) showed rankings shift under contamination-free reformulation. Our variance decomposition synthesizes these findings.

Evaluation methodology. Weidinger et al. (2025) provides theoretical foundations for triangulated measurement but not implementation. We operationalize these with concrete metrics (RR), falsifiable thresholds, and cost-stratified protocols (Table 5).

Dynamic benchmarks. Dynamic benchmarks (LiveBench, LiveCodeBench) address contamination; our claim-type distinction specifies *which* inferences each supports.

Test-time compute. Zhao et al. (2025a) analyze reasoning scaling. We hypothesize test-time compute does not resolve interpolation-extrapolation confounds (Appendix I).

See Appendix K for extended discussion.

8. Conclusion

Core thesis. Rankings reliably compare methods; capability claims require scope evidence. Benchmarks can be simultaneously reliable for method comparison and invalid for capability inference.

Summary of evidence. Rankings remain stable under structure-preserving transformations (empirically $\tau > 0.9$, exceeding our $\tau > 0.8$ threshold), while absolute scores collapse under structural novelty—degradations of 95% (ARC-AGI) and 67% (SWE-bench Pro). The variance decomposition attributes these effects to interpolation, extrapolation, gaming, and noise components.

Call to action. We invite the community to: (1) validate the Robustness Ratio on MMLU (protocol in Appendix N); (2) construct GSM8K-Novel to test whether LLM rankings transfer as robustly as vision rankings; (3) adopt claim-type labeling (*[Method]*, *[Interpolation]*, *[Extrapolation]*) in publications.

Limitations. The decomposition is operational, not metaphysical. Validity protocols will themselves become optimization targets—we view evaluation as an ongoing adversarial game, not a solvable problem.

Closing. Our framework makes testable predictions: rankings should transfer when task structure is preserved ($\tau > 0.8$); capability claims require transfer evidence. If structure-preserving variants show $\tau < 0.5$, or matched-score models show equivalent deployment performance regardless of training composition, our framework is wrong. We invite the community to test it.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6430–6439, 2019.

Alaa, A. et al. Medical llm benchmarks should prioritize construct validity. In *International Conference on Machine Learning (Position Paper)*, 2025.

Alvarez-Melis, D. and Fusi, N. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21428–21439, 2020.

ARC Prize Foundation. Arc-agi-2 benchmark and competition. https://arcprize.org, 2025a. Accessed January 2026.

ARC Prize Foundation. Arc-agi-3: The first interactive reasoning benchmark for ai agents. https://arcprize.org/arc-agi/3/, 2025b. July 2025.

Bandel, E., Aharonov, R., Shmueli-Scheuer, M., Shnayderman, I., Slonim, N., and Ein-Dor, L. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 596–609, 2022.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Bordt, S., Srinivas, S., Boreiko, V., and von Luxburg, U. How much can we forget about data contamination? In *International Conference on Machine Learning*, 2025.

Burnell, R. et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138, 2023.

Chen, Y. et al. Reliabilitybench: Benchmarking agent reliability under perturbations. *arXiv preprint arXiv:2601.06112*, 2025.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. Investigating data contamination in modern benchmarks for large language models. In *NAACL*, 2024.

Domínguez-Olmedo, R., Dorner, F. E., and Hardt, M. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations*, 2025. Oral presentation.

Donoho, D. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):731–768, 2017.

Epoch AI. Independent evaluation of o3 on frontiermath. https://techcrunch.com/2025/04/20/openais-o3-ai-model-scores-lower-on-a-benchmark-than-the-company-initially-implied/, 2025. Accessed January 2026.

Errica, F. et al. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. *arXiv preprint arXiv:2406.12334*, 2024.

Fan, Y., Zhang, Y., et al. Math-perturb: Evaluating llms' math reasoning through perturbation-based benchmarks. In *ICML*, 2025. arXiv:2502.06453.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021.

Hsu, K. et al. Compositional-arc: Assessing systematic generalization in abstract spatial reasoning. *arXiv preprint arXiv:2504.01445*, 2025.

Hu, L., Li, Q., Xie, A., Jiang, N., Stoica, I., Jin, H., and Zhang, H. Gamearena: Evaluating llm reasoning in competitive game environments. In *ICLR*, 2025.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Livecodebench: Holistic and contamination free evaluation of large language models for code. https://livecodebench.github.io, 2024. Accessed January 2026.

Kane, M. T. Validation. In Brennan, R. L. (ed.), *Educational Measurement*, pp. 17–64. American Council on Education/Praeger, 4th edition, 2006.

Koch, L. M., Baumgartner, C. F., and Berens, P. Distribution shift detection for the postmarket surveillance of medical AI algorithms. *npj Digital Medicine*, 7(1):120, 2024.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

Liao, T., Taori, R., Raji, I. D., and Schmidt, L. Are we learning yet? a meta review of evaluation failures across machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Lunardi, R., Della Mea, V., Mizzaro, S., and Roitero, K. On robustness and reliability of benchmark-based evaluation of llms. *arXiv preprint arXiv:2509.04013*, 2025.

Messick, S. Validity. In Linn, R. L. (ed.), *Educational Measurement*, pp. 13–103. Macmillan, 3rd edition, 1989.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. *International Conference on Machine Learning*, pp. 7721–7735, 2021.

Miserendino, S. et al. SWE-Lancer: Can frontier LLMs earn money on freelance coding platforms? *arXiv preprint arXiv:2502.12115*, 2025.

Nalbandyan, A. et al. Score: Systematic consistency and robustness evaluation for large language models. *arXiv preprint arXiv:2503.00137*, 2025. NAACL 2025.

Needham, C. et al. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025.

Ni, A. et al. L2ceval: Evaluating language-to-code generation capabilities of large language models. In *TACL*, 2024.

Nunnally, J. C. and Bernstein, I. H. *Psychometric Theory*. McGraw-Hill, 3rd edition, 1994.

Qian, X. et al. Benchmark2: Systematic evaluation of llm benchmarks. *arXiv preprint arXiv:2601.03986*, 2025.

Rabanser, S., Günnemann, S., and Lipton, Z. C. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Salaudeen, O. and Hardt, M. Imagenot: A contrast with imagenet preserves model rankings. *arXiv preprint arXiv:2404.02112*, 2024.

Scale AI. Swe-bench pro: A contamination-resistant benchmark for software engineering agents. https://scale.com/leaderboard/swe_bench_pro_public, 2025. Accessed January 2026.

Sherborne, T. et al. Measuring what matters: Construct validity in llm benchmarks. In *Advances in Neural Information Processing Systems*, 2025.

Shi, W. et al. Quantifying training data overlap in large language models. *arXiv preprint*, 2025. Training data contamination analysis.

Sweller, J. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

Tao, T., Zhang, Y., and Liu, W. Magis: Benchmarking agentic coding with rigorous test suites. In *arXiv preprint arXiv:2402.11161*, 2024.

Teney, D., Lin, Y., Oh, S. J., and Abbasnejad, E. Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems*, 36, 2023.

Thomas, D., Shang, Y., Tian, J., et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

Wang, Y. and Zhao, X. Evaluating llm robustness to adversarial textual perturbations on reasoning datasets. *arXiv preprint*, 2024.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Weidinger, L. et al. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*, 2025. DeepMind.

Westrick, P. A., Marini, J. P., Young, L., Ng, H., Shmueli, D., and Shaw, E. J. Validity of the SAT for predicting first-year grades and retention to the second year. Research report, College Board, 2019.

White, C., Dooley, S., Roberts, M., Pal, A., Feber, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saaber, K., Keutzer, K., et al. Livebench: A challenging, contamination-free llm benchmark. https://livebench.ai, 2024. Accessed January 2026.

Xu, C., Yan, N., Guan, S., Jin, C., Mei, Y., Guo, Y., and Kechadi, T. Dcr: Quantifying data contamination in llms evaluation. In *EMNLP*, 2025. arXiv:2507.11405.

Yang, H. et al. Detecting data contamination from reinforcement learning post-training for large language models. *arXiv preprint arXiv:2510.09259*, 2025. Introduces Self-Critique method for RL-trained model contamination detection.

Yao, S., Narasimhan, K., and Wen, Q. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *arXiv preprint arXiv:2406.12045*, 2024.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Duan, N., Zhang, X., et al. A careful examination of large language model performance on grade school arithmetic. In *NeurIPS*, 2024. arXiv:2405.00332.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Zhao, J. et al. Genprm: Scaling test-time compute of process reward models via generative verifiers. *arXiv preprint arXiv:2504.00891*, 2025a.

Zhao, Q., Huang, Y., Lv, T., Cui, L., Sun, Q., Mao, S., Zhang, X., Xin, Y., Yin, Q., Li, S., and Wei, F. Mmlu-cf: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13371–13391, 2025b.

Zhu, K. et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint*, 2024.

A. Mathematical Proofs

This appendix provides formal statements and proofs for the propositions referenced in the main text.

Proposition A.1 (Ranking Stability)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing. For $\tilde{s}(h) = g(s(h))$: $s(h_A) > s(h_B) \iff \tilde{s}(h_A) > \tilde{s}(h_B)$*

Proposition A.2 (Score Contraction Under Symmetric Noise)**.** *For K-class classification with symmetric label-flip rate η (where each true label is independently replaced with a uniformly random incorrect label with probability η): $\mathrm{Acc}_\eta(h) = \alpha \cdot \mathrm{Acc}_0(h) + \beta$ where $\alpha = 1 - \eta - \frac{\eta}{K-1} > 0$ for $\eta < \frac{K-1}{K}$ and $\beta = \frac{\eta}{K-1}$. Rankings preserved; absolute scores contract toward $\frac{1}{K}$.*

A.1. Proof of Proposition A.1 (Ranking Stability)

Proof. Let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing. By definition of strict monotonicity, $a > b \iff g(a) > g(b)$. Substituting $a = s(h_A)$ and $b = s(h_B)$:

$$s(h_A) > s(h_B) \iff g(s(h_A)) > g(s(h_B)) \iff \tilde{s}(h_A) > \tilde{s}(h_B)$$

Thus rankings are preserved under any strictly increasing transformation. □

A.2. Proof of Proposition A.2 (Score Contraction)

Proof. For K-class classification with clean accuracy $\mathrm{Acc}_0(h)$ and symmetric label-flip rate η:

$$\begin{aligned} \mathrm{Acc}_\eta(h) &= (1-\eta) \cdot \mathrm{Acc}_0(h) + \eta \cdot \frac{1 - \mathrm{Acc}_0(h)}{K-1} \\ &= \underbrace{\left(1 - \eta - \frac{\eta}{K-1}\right)}_{\alpha} \mathrm{Acc}_0(h) + \underbrace{\frac{\eta}{K-1}}_{\beta} \end{aligned} \tag{1}$$

For $\eta < \frac{K-1}{K}$, we have $\alpha > 0$, ensuring the transformation is strictly increasing. Rankings are preserved while absolute scores contract toward β. □

B. Extended Evidence Discussion

This section elaborates on the empirical findings summarized in the main text.

B.1. The Fine-Tuning Paradox

Fine-tuning is training, so it changes the method per the definition in Section 2. We resolve this apparent paradox by decomposing "method" into two components:

1. *Core method*: architecture, optimization algorithm, and primary training data
2. *Exposure component*: incidental benchmark overlap from data curation

Fine-tuning equalization controls for exposure while preserving core method—analogous to controlling for confounds in clinical trials. Post-equalization gaps measure: "Given equal exposure, which core method produces better performance?" The need for such adjustment supports the view that raw scores measure a mixture of constructs.

B.2. Contamination Detection Failures

Detection accuracy varies dramatically by training paradigm. The following estimates are *extrapolated qualitatively* from Domínguez-Olmedo et al. (2025), who report that contamination detection achieves high accuracy for standard fine-tuned models but fails for reasoning models. These figures are illustrative estimates, not precise measurements:

Training Type	Detection Acc.*	Implication
SFT only	>90%	Detectable
RL training	∼50%	Random guessing
SFT + CoT	∼50%	Concealed signals

*Illustrative estimates; not empirically measured.

For reasoning models (o1, o3, DeepSeek-R1), contamination may be undetectable via traditional methods, unquantifiable, and therefore unadjustable. A plausible explanation: GRPO (Group Relative Policy Optimization) and PPO (Proximal Policy Optimization) training use importance sampling and clipping objectives that may reduce the effectiveness of log-probability-based detection methods—these techniques can transform output distributions in ways that obscure memorization signals. Additionally, reasoning models may internalize reasoning processes rather than memorizing specific sequences.

Emerging detection methods for reasoning models. Recent work suggests partial solutions (Table 6).

Table 6. Emerging contamination detection methods for reasoning models. Improvement figures are reported values from recent preprints (2024–2025); independent validation pending.

Method	Improvement	Mechanism
Self-Critique (Yang et al., 2025)	+30% AUC (reported)	Detects entropy collapse in RL-trained models
Contamination Detection via Distributional Shift (CDD)	+21–30% rel. (reported)	Black-box detection via output distribution peakedness
Perturbation testing	>15% drop signal	Behavioral validation via systematic input modification
Watermarking	Verifiable	Proactive cryptographic markers (new benchmarks only)

Recommended detection pipeline. For reasoning model evaluation:

1. Apply CDD for output distribution analysis
2. Apply perturbation testing for behavioral validation
3. Cross-validate with temporal membership inference if training dates are known
4. Prefer dynamic benchmarks (LiveBench, LiveCodeBench) that avoid contamination by construction

Note: To our knowledge, no single method currently achieves >80% accuracy on RL-trained models; ensemble approaches may be required.

B.3. Application-Layer Optimization

The "refinement loop" paradigm suggests that benchmark performance increasingly depends on *how models are used* rather than *how they were trained* (Table 7).

Table 7. ARC-AGI-2 performance across systems (ARC Prize Foundation, 2025a). Results span 4%–75% on public set, demonstrating high variance from methodology.

System	Public	Semi-Priv.	Method
Baseline (o3)	4%	—	Single-pass
Giotto.ai	22%	—	Deep reasoning
Extended reasoning model	45%	—	Extended inference
Poetiq (refinement)	75%	54%	Refinement loop
Human average	60%	60%†	—

†Human average is 60% across all evaluation sets (calibrated to same difficulty).

Results span 4%–75% on public ARC-AGI-2, demonstrating high variance from methodology alone. The public/semi-private gap for AI systems (Poetiq: 75% vs. 54%) suggests public sets may have leaked or contain easier problems, while human

performance remains consistent at 60% across sets. This reinforces our core point: benchmark scores reflect procedures, not fixed artifact properties.

B.4. Vision-to-Language Extrapolation and Direct Tests

Why the distinction matters. The ImageNot result relies on physical regularities in images; language presents a different case where regularities are cultural patterns in human communication. The difference in regularity types has implications for how we interpret ranking transfer evidence:

- **Vision**: Rankings transfer to *genuinely novel distributions* (ImageNot uses different images, not paraphrased versions). This suggests vision model rankings reflect architecture/training quality rather than dataset-specific fitting.
- **Language**: Rankings are stable under *linguistic perturbations* (paraphrase, format changes) but sensitive to *methodological changes* (prompt templates, evaluation protocols). Thomas et al. (2024) document up to 8 position shifts from prompt template changes alone.

What would close the gap? GSM8K-Novel—a math benchmark with different surface statistics but preserved mathematical concepts—would test whether LLM rankings transfer as robustly as vision rankings (author estimate: $\sim 50\times$ baseline evaluation cost). If rankings transfer, the vision-LLM gap closes. If rankings fail, it confirms that LLM benchmark scores are more distribution-specific than vision benchmarks.

Evaluation-aware behavior. Reports from frontier model evaluations suggest evaluation-aware behavior: some models reportedly reproduce benchmark canary strings and exhibit evaluation-specific patterns (Burnell et al., 2023), suggesting Var_{game} is a real phenomenon. Benchmarks both overestimate (contamination) and underestimate (single-pass misses latent capability); the direction depends on deployment context.

B.5. Contamination Effect Estimates

Table 8 summarizes empirical contamination effect estimates from the literature. These estimates come from different methodologies and are not directly comparable, but collectively establish that contamination effects are substantial; they do not provide precise point estimates.

Table 8. Contamination and distribution shift effects by benchmark.

Benchmark	Metric	Effect	Source
MMLU	Post-Nov-2023 boost	+7 pts	Domínguez-Olmedo et al. (2025)
GSM8K	Post-Nov-2023 boost	+17 pts	Domínguez-Olmedo et al. (2025)
MMLU	Prompt sensitivity	up to 10% variation	Wang et al. (2024)
MMLU-CF	Contam-free vs. original	-14.6 pts (88%→73%)	Zhao et al. (2025b)
MMLU-CF	Ranking changes	"Changed considerably"	Zhao et al. (2025b)
SWE-bench	Verified → Pro	70% → 23%	Scale AI (2025)
SWE-bench	Pro pub. → comm.	23% → 15–18%	Scale AI (2025)
Paraphrase	Answer variance	15–30% differ	Lunardi et al. (2025)
Paraphrase	Ranking stability	$\tau > 0.9$ preserved	Lunardi et al. (2025)

C. Validity Protocol Details

This section specifies concrete thresholds and procedures for validity assessment.

C.1. Transfer Test Thresholds

We propose three confidence levels based on transfer performance (Table 9). These thresholds are author-proposed guidelines inspired by psychometric reliability standards, not direct derivations from prior work.

Table 9. Transfer test thresholds for capability claims (author-proposed, inspired by psychometric reliability standards).

Confidence	Criterion
High	$\geq$90% of benchmark score on $\geq$3 held-out distributions; variance $<$10%
Moderate	$\geq$75% of benchmark score on $\geq$2 held-out distributions
Preliminary	Significant positive transfer ($p < 0.05$) to $\geq$1 distribution

C.2. Dynamic Benchmark Protocols

For continuously updated benchmarks (e.g., SWE-bench-Live, LiveCodeBench), temporal analysis is essential (Table 10).

Table 10. Evaluation protocols for dynamic benchmarks.

Protocol	Requirement	Rationale
Longitudinal comparison	Rankings stable across monthly updates; report evaluation date with all scores	Enables temporal analysis and reproducibility
Freshness decay	Define acceptable decay rate (e.g., $\leq$5% per quarter)	Faster decay suggests training-distribution dependence

C.3. Reasoning Model Protocols

For models using test-time compute (o1, o3, DeepSeek-R1), standard evaluation is insufficient. Table 11 specifies additional requirements.

Table 11. Evaluation protocols for reasoning models with test-time compute.

Protocol	Requirement	Rationale
Cost-normalized metrics	Report accuracy/$ and accuracy/FLOP alongside raw scores	Enables fair comparison across compute levels
Multi-level evaluation	Report at low ($\leq$1K tokens), medium ($\leq$10K), and high ($\leq$100K) compute budgets	Detects whether rankings change with compute
Elicitation vs. search	Test on truly novel problems; analyze reasoning trace coherence	Distinguishes learned strategies from brute-force search

D. Psychometric Foundations

Understanding the validity protocols above requires situating ML benchmarks within the broader context of measurement theory. ML benchmarks lack the validation infrastructure that psychometric tests have accumulated over a century (Table 12). This asymmetry explains why we advocate epistemic humility for ML capability claims.

The asymmetry in our conclusions reflects asymmetry in accumulated evidence, not fundamental measurement differences. When ML accumulates comparable validation infrastructure, capability claims will be similarly warranted.

The SAT as bounded predictive validity. SAT scores predict college GPA with moderate validity ($r \approx 0.48$–0.51 for SAT alone; $r \approx 0.60$ combined with HSGPA; Westrick et al. 2019), but this validity is bounded. The SAT does not predict professional domain performance—nor should it. This is not a validity failure; it is bounded scope, and SAT documentation reflects this. ML benchmarks face analogous limitations, but unlike the SAT, these bounds are rarely specified.

Messick's (1989) unified validity theory provides the theoretical foundation: "validity is not a property of the test or assessment as such, but rather of the meaning of the test scores" (Messick, 1989). Validity is established by accumulating

Table 12. Validation infrastructure: psychometrics vs. ML evaluation.

Validation Type	Psychometrics	ML Benchmarks
Predictive validity	IQ correlates with outcomes ($r \approx 0.2$–0.5) for decades[†]	No comparable long-horizon studies
Theoretical grounding	Factor analysis identifies g; neuroimaging correlates	No construct specification for "reasoning"
Irrelevant variance	Coaching effects characterized (0.1–0.3 SD)	No systematic contamination auditing
Measurement invariance	Tests verified across demographic groups	Assumed equivalent across architectures

[†] Nunnally & Bernstein (1994); see also meta-analyses on cognitive ability and job performance.

evidence that score interpretations are warranted for specific uses. ML benchmarks have accumulated evidence for method-ranking interpretations but not for capability interpretations.

E. The Variance Decomposition Framework

Building on the psychometric principles above, we now formalize the variance decomposition framework. This section formalizes the theoretical framework underlying our position.

E.1. Variance Decomposition

We propose that benchmark score variance can be partitioned into operationally distinct components (Table 13):

$$\text{Var}(\text{Score}) = \text{Var}_{\text{interp}} + \text{Var}_{\text{extrap}} + \text{Var}_{\text{game}} + \text{Var}_{\text{noise}} \tag{2}$$

Table 13. Variance decomposition components.

Component	Estimate	Interpretation
$\text{Var}_{\text{interp}}$	$\hat{\alpha}$	Variance explained by training-test distribution overlap
$\text{Var}_{\text{extrap}}$	$\hat{\beta}$	Variance explained by transfer to structurally novel tasks
Var_{game}	$\hat{\gamma}$	Variance from models actively optimizing for evaluation patterns
$\text{Var}_{\text{noise}}$	—	Measurement noise, prompt sensitivity, random seeds

Operational estimation. For any benchmark, these components can be empirically estimated via regression. We hypothesize that static benchmarks (MMLU, GSM8K) have high $\hat{\alpha} + \hat{\gamma}$ while interactive benchmarks (ARC-AGI-3; preliminary) have higher $\hat{\beta}$.

Caveat. This decomposition is operational rather than metaphysical. Even if constructs are entangled, the epistemological distinction—different claims require different evidence—holds regardless of ontological separability.

Practical limitation. Computing $\hat{\alpha}$ requires contamination estimates for each model-benchmark pair. For closed models, training data composition is unavailable; for open models, contamination detection fails for reasoning models ($\sim$50% accuracy). We rely on indirect evidence as proxies: score drops under reformulation (MMLU $\rightarrow$ MMLU-CF: 14–16 points), distribution shift effects (SWE-bench Verified $\rightarrow$ Pro: 47+ points).

F. Asymmetric Contamination Analysis

Proposition A.2 (ranking preservation under symmetric noise) does not apply when contamination is asymmetric—when some models have seen the original test set and others have not.

F.1. Mechanism

Consider two models with equal underlying capability but different training exposure (Table 14).

This ranking reversal illustrates how contamination can inflate scores on original benchmarks while failing to transfer to semantically equivalent reformulations.

Table 14. Asymmetric contamination: ranking reversal example.

Model	Training Exposure	Original MMLU	Paraphrased	Mechanism
A	Contaminated (saw questions)	85%	70% (↓15%)	Memorization-dependent
B	Uncontaminated	85%	83% (↓2%)	Generalization-based
Ranking		A = B	B > A	*Rankings flip*

F.2. Evidence

MMLU-CF (Zhao et al., 2025b) provides direct evidence. By constructing contamination-free reformulations of MMLU questions, the authors found:

- "Performances significantly dropped"—consistent with contamination inflation on original.
- "Rankings changed considerably"—inconsistent with universal ranking preservation.

This is precisely the asymmetric contamination effect: models with high exposure to original MMLU lose their advantage on reformulated versions.

F.3. Implications

Our validity protocols are most important precisely when asymmetric contamination is suspected. Paraphrase testing (the Robustness Ratio) distinguishes contaminated from uncontaminated high performers. If a model's score drops substantially on paraphrased versions, this suggests surface-form memorization rather than robust understanding.

For cross-cohort comparisons (e.g., models released before vs. after MMLU became widely used in training), rankings on the original benchmark are unreliable. Paraphrased or contamination-free versions provide cleaner capability estimates.

G. Extended Introduction

G.1. Epistemic Transparency

We are explicit about confidence levels:

Claim	Confidence	Evidence
Rankings transfer (vision)	High	ImageNot
Rankings transfer (LLM)	Moderate	Lunardi et al.
Contamination inflates	High	Domínguez
TTC: search vs. extrapolation	Uncertain	Ongoing
Var_{game} exists	Moderate	Frontier model cases

G.2. Framework Scope

Our framework applies with varying force: strongly for ill-defined constructs (reasoning, understanding); weakly for verifiable benchmarks (code, proofs) where ground truth via execution provides stronger validity; and with limited application for saturated benchmarks (>90%) where interpolation may be equalized at frontier. See Appendix J for detailed scope discussion.

H. Extended Theoretical Discussion

H.1. Resolving Entanglement via Causal Decomposition

A stronger objection: if better training data produces more capable models, then measuring "data quality" *is* measuring capability. We resolve this via causal structure:

$$\text{Data Quality} \rightarrow \text{Interpolation} \rightarrow \text{Benchmark Score}$$

Data Quality → Extrapolation Capability → Deployment Performance

The path through Interpolation is short and measurable via scaling law residuals; the path through Extrapolation requires longitudinal transfer tests. Our distinction is *epistemological*: these paths yield different evidence requirements. The variance decomposition operationalizes this: $\hat{\alpha}$ captures the interpolation path, $\hat{\beta}$ captures the extrapolation path.

H.2. Falsifiability

Our position would be refuted if: (1) benchmark scores predicted deployment performance equally regardless of distribution similarity; (2) fine-tuning equalization failed to reduce score gaps; (3) matched-score models showed equivalent held-out performance regardless of training data composition.

I. Test-Time Compute Analysis

This appendix provides detailed analysis of reasoning models and test-time compute.

I.1. The Classification Problem

When R1 solves a math problem by generating a long reasoning chain—is this interpolation (optimized RL objective) or extrapolation (genuine reasoning that transfers)? We distinguish training-time interpolation (minimizing loss given training compute) from inference-time interpolation (converting inference compute into correct answers on familiar structures).

I.2. Current Evidence

The ARC-AGI collapse (75.7% → 3–4%) discussed in Section 3 persists despite extended reasoning. (Poetiq's 54% uses specialized refinement methods.) The ARC-AGI-3 preview (January 2026; preliminary, benchmark not finalized) is consistent with this: the best AI agent achieves only 12.6% on interactive mini-games where human performance approaches ceiling (ARC Prize Foundation, 2025b). This provides preliminary support for our framework: ARC-AGI-3's interactivity increases $\hat{\beta}$, exposing the gap between interpolation and extrapolation.

I.3. Reconciling ImageNot with ARC-AGI

A potential contradiction: ImageNot shows $\tau > 0.9$ ranking transfer, but ARC-AGI shows catastrophic drop (75.7%→3–4%). We propose a reconciliation via the **Task Structure Preservation Principle**: Rankings transfer when task structure is preserved (ImageNot: same domain, different images); rankings fail when structure is deliberately altered (ARC-AGI-2: redesigned solution patterns). Under our framework, this is what separates interpolation (structure-dependent) from extrapolation (structure-independent).

J. Scope, Adoption, and Future Directions

This appendix discusses scope of critique, adoption incentives, and future research directions.

J.1. Scope of Our Critique

Our critique applies with varying force depending on deployment context:

Critique applies strongly when: (a) Deployment distribution differs substantially from benchmark distribution—e.g., deploying a coding assistant trained on popular repositories to proprietary enterprise codebases; (b) Capability claims are made from benchmark scores—e.g., "Model X demonstrates broad knowledge" based on MMLU performance; (c) Safety-critical decisions are based on benchmark performance—e.g., using benchmark scores to certify medical AI systems.

Critique applies less strongly when: (a) Deployment distribution matches benchmark distribution—e.g., customer service bots answering FAQ-style questions similar to training data; (b) Scores are used only for ranking methods within similar conditions—our thesis explicitly endorses this use; (c) Tasks have verifiable outputs and comprehensive test coverage—though validity concerns persist.

For narrow deployments on benchmark-similar tasks, interpolation performance *is* the product. High MMLU correlates

with better FAQ performance precisely because task structure is preserved. Our critique targets the unwarranted leap from "performs well on benchmark" to "has general capability."

J.2. Robustness Ratio Threshold Hypotheses

We propose $RR < 1.5$ as indicating robust performance and $RR > 3$ as indicating fragile pattern-matching. These thresholds are *preliminary hypotheses* requiring validation: (H1) Models with low RR should show better OOD transfer; (H2) Reasoning models (o1, R1) should have lower RR than base models; (H3) RR should correlate with deployment performance on novel tasks. We invite the community to test these hypotheses. Computing RR requires approximately 1–2 hours of API queries per benchmark (author estimate).

Validation pathway. Table 15 specifies the steps to test H1–H3. We have not conducted this validation; the framework is ready for community testing.

Table 15. Validation pathway for RR hypotheses.

Step	Method	Tests
Data collection	Compute RR for 20+ models on 500 MMLU questions (5 paraphrase variants each)	—
Correlation test	Correlate RR with MMLU→MMLU-Pro degradation	H1
Reasoning test	Compare RR for base vs. reasoning variants (e.g., GPT-4o vs. o3)	H2
Estimated cost: ∼1.5–2× baseline API budget; ∼40 hours researcher time		

J.3. Goodhart Dynamics and Endgame

Once validity protocols become standard, they become optimization targets—our proposals face the same dynamics we diagnose. We take a position on the endgame: *evaluation is an adversarial game requiring continuous innovation*. We believe there is no permanent solution; the goal is not to "solve" evaluation but to maintain an honest assessment of what benchmarks can and cannot tell us. Arms race measures may buy time for the field to develop better tools. Benchmark-based evaluation may ultimately give way to deployment-based evaluation for capability claims, with benchmarks reserved for method comparison where they excel.

J.4. Emerging Deployment-Focused Evaluation

This shift is already underway. Recent benchmarks directly assess economic and deployment value rather than isolated task performance: SWE-Lancer (Miserendino et al., 2025) evaluates agents on real freelance coding tasks with monetary stakes; agent benchmarks assess multi-step workflows in production environments. These approaches align with our recommendation to ground capability claims in deployment performance rather than benchmark proxies.

The deployment evaluation endgame. Validity protocols will themselves become optimization targets. One possible endgame is deployment-based evaluation for capability claims (Table 16). Benchmarks would remain useful for method comparison (their strength) but would be replaced by deployment studies for capability claims (their weakness). This shifts evaluation cost but provides ground truth unavailable from offline benchmarks.

J.5. Multimodal Models

For multimodal models (GPT-4V, Gemini, Claude 3), our framework applies with additional complexity: capability claims may involve cross-modal transfer (vision-to-language reasoning) where interpolation/extrapolation boundaries are less clear. Preliminary evidence suggests multimodal benchmarks exhibit similar saturation patterns (Yue et al., 2024), and our validity protocols extend naturally: paraphrase robustness for text components, augmentation robustness for visual components, and compositional testing for cross-modal reasoning.

Table 16. Deployment-based evaluation approaches.

Approach	Method	Ground Truth
SWE-Lancer	Real freelance coding tasks with monetary stakes	Economic value
Agent benchmarks	Multi-step workflows in production environments	Task completion
A/B deployment	Field evaluation with business metrics	User outcomes

J.6. Priority Research Directions

Table 17 ranks key validation experiments by priority and estimated cost.

Table 17. Priority research directions for framework validation.

Priority	Experiment	Goal	Cost
1	GSM8K-Novel	Test LLM ranking transfer to structurally novel math problems	$\sim 50\times$ baseline
2	ARC-AGI-3 evaluation	Validate interactive reasoning (preliminary; benchmark not final)	Benchmark-dependent
3	Contamination detection	Develop reliable detection for reasoning models (o1, o3, R1)	Research effort
4	Interp./extrap. estimation	Validate $\hat{\alpha}$, $\hat{\beta}$ via independent metrics	$\sim 2\times$ baseline

K. Additional Related Work

Our framework builds on several research threads: (1) **Construct validity**: "Measuring What Matters" (Sherborne et al., 2025) provides a taxonomy of validity failures; Alaa et al. (2025) show high benchmark scores do not predict clinical performance for medical LLMs. (2) **Contamination-free evaluation**: MMLU-CF (Zhao et al., 2025b) demonstrates "rankings changed considerably" on contamination-free reformulations, directly supporting our interpolation-dominated interpretation. (3) **Contamination dynamics**: Bordt et al. (2025) show contamination can be "forgotten" for over-trained models, suggesting effects depend on training dynamics. (4) **Test-time compute**: Process reward model scaling (Zhao et al., 2025a) demonstrates test-time compute improves performance but doesn't address structural transfer.

L. Task Structure Operationalization

This appendix grounds Definition 2.6's thresholds in established literature. Table 18 provides a quick reference; detailed justifications follow.

Table 18. Task structure preservation thresholds (Definition 2.6).

Metric	Threshold	Interpretation	Source
Vocabulary overlap	Jaccard > 0.7 or PAD < 0.5	Domains sufficiently similar	Domain adaptation (Ben-David et al., 2006)
Reasoning depth	Ratio $\in [0.5, 2.0]$	Approximate complexity preservation	Cognitive load theory (threshold proposed)
Ranking correlation	$\tau > 0.8$	Strong reliability	Psychometrics (Nunnally & Bernstein, 1994)

L.1. Threshold Justification

Vocabulary overlap (>0.7 Jaccard or PAD < 0.5). The Proxy A-Distance (PAD) from domain adaptation theory (Ben-David et al., 2006) provides a theoretically-grounded measure of distribution similarity. We *propose* PAD < 0.5 as indicating domains "sufficiently similar" for transfer and PAD > 1.0 as indicating clearly distinguishable domains; these thresholds are derived from empirical observations in the domain adaptation literature but require validation in the ML benchmark context. The Jaccard > 0.7 correspondence is our proposed heuristic.

Reasoning depth ratio ($\in [0.5, 2.0]$). Cognitive load theory (Sweller, 1988) establishes that reasoning difficulty scales with "element interactivity"—the number of elements that must be processed simultaneously. We propose that a $2\times$ change in reasoning depth corresponds to approximately one standard deviation in cognitive load (author estimate based on cognitive load literature), beyond which qualitatively different strategies may be required.

Operationalizing reasoning depth. We acknowledge that "reasoning depth" risks subjectivity. To address this, we propose three complementary operationalizations:

1. **Reference solution length ratio**: For tasks with known solutions, compute $\text{depth}_{\text{ratio}} = \text{tokens}(B')/\text{tokens}(B)$ using minimal correct solutions. This is objective but requires reference solutions.
2. **CoT token ratio**: For chain-of-thought models, compute the ratio of generated reasoning tokens. If model M produces n tokens on benchmark B and n' on variant B', then $\text{depth}_{\text{ratio}} \approx n'/n$. This assumes CoT length correlates with reasoning complexity—an assumption supported by chain-of-thought scaling work (Wei et al., 2022) but domain-dependent in practice.
3. **Reasoning tree complexity**: Convert sequential reasoning chains to tree structures following recent chain-of-thought analysis methods and compare: (a) tree depth, (b) branching factor, (c) number of distinct reasoning steps. Ratios outside $[0.5, 2.0]$ on any metric suggest structure alteration (thresholds proposed).

For practical use, we recommend the CoT token ratio as the default (cheapest to compute) with reasoning tree analysis for high-stakes decisions. When metrics disagree, the most conservative (largest deviation from 1.0) should be used.

Inter-annotator agreement. A limitation of criteria 2 and 4 (reasoning depth and solution template) is their dependence on researcher judgment about "intended solution steps" and "solution template." Future work should establish inter-annotator reliability for these criteria, targeting Cohen's $\kappa > 0.7$ before treating them as objective. In the absence of formal annotation, post-hoc ranking correlation (τ) serves as a model-independent tiebreaker: where annotators disagree on structure preservation, empirical τ resolves the disagreement without relying on any single annotator's judgment. Our falsifiability claim rests on the combined criterion-plus-τ system—criteria specify structure preservation *a priori*; τ validates it *post hoc*.

Ranking correlation ($\tau > 0.8$). The 0.8 threshold is adapted from psychometric reliability standards (Nunnally & Bernstein, 1994), which establishes 0.8 as the threshold for adequate internal consistency (Cronbach's α). We extend this principle to rank correlation: $\tau > 0.8$ indicates strong ranking stability suitable for individual-level decisions; $\tau \in [0.6, 0.8]$ indicates moderate stability suitable for group comparisons; $\tau < 0.6$ indicates weak stability requiring caution. This adaptation is a proposed heuristic, not a direct derivation.

L.2. Alternative Metrics

Beyond the thresholds in Definition 2.6, researchers may consider complementary metrics (Table 19). For high-stakes decisions, multiple metrics should converge.

M. Distribution Shift Taxonomy and Diagnostics

Practitioners must distinguish three cases when evaluating performance on new data (Table 20).

Ex ante diagnostic protocol. Before observing transfer:

1. *Structural criteria*: Check a priori criteria from Definition 2.6. If satisfied, predict (a) or (b); if violated, predict (c).
2. *Embedding distance*: Compute mean cosine distance of new examples from benchmark examples using a general-purpose encoder (e.g., sentence transformers). Novelty $> 2\sigma$ from mean suggests (a) or (c); moderate novelty suggests (b).
3. *Difficulty calibration*: If degradation correlates with difficulty metrics (reasoning steps, vocabulary complexity), likely

Table 19. Alternative metrics for task structure assessment.

Metric	Method	Threshold	Interpretation
Optimal Transport Dataset Distance (OTDD)	Optimal transport between task embeddings (Alvarez-Melis & Fusi, 2020)	< 0.3 (proposed)	Structure preserved
Task2Vec	Cosine similarity of Fisher information embeddings (Achille et al., 2019)	> 0.8 (proposed)	Similar task structure
Measurement invariance	Confirmatory factor analysis (CFA) across populations	Standard fit indices	Same latent construct

Table 20. Distribution shift taxonomy and diagnostic indicators.

Case	Characteristic	Diagnostic Signal	Implication
(a) Tail of same dist.	Rare but valid instances of same construct	Degradation correlates with difficulty metrics	Benchmark scope still applies
(b) Different slice	Within support but jagged capabilities (Ni et al., 2024)	Degradation is slice-specific (domains/formats)	Slice-specific remediation
(c) Different construct	New benchmark measures distinct property	$\tau < 0.5$ between benchmarks	Cross-benchmark comparison invalid

(a). If degradation is slice-specific (certain domains/formats), likely (b).

These are heuristics with uncertain reliability. Developing validated detection methods is a research direction we identify in Section 5. Post hoc, ranking correlation (τ) provides definitive diagnosis.

N. Robustness Ratio Validation

Threshold derivation with independent corroboration. Initial thresholds derive from Lunardi et al. (2025): 15–30% answer variance under paraphrase versus <5% seed variance yields $RR_{\text{fragile}} \approx 3.0$–$6.0$. Table 21 summarizes independent corroboration.

Table 21. Independent corroboration of RR thresholds.

Source	Benchmark	Finding	Implied RR
Lunardi et al. (2025)	Various	15–30% paraphrase var. vs. <5% seed	3.0–6.0
Zhu et al. (2024)	PromptBench	15–30% drops under semantic paraphrase	3.0–6.0
Nalbandyan et al. (2025)	MMLU-Pro	Up to 10% accuracy fluctuation	∼2.0
Chen et al. (2025)	ReliabilityBench	8.8% decline at $\epsilon = 0.2$	∼1.8

These converging estimates support $RR > 3$ as indicating fragile pattern-matching.

Additional validation comes from: (1) Errica et al. (2024) introduce complementary sensitivity/consistency metrics showing similar variance patterns; (2) Wang & Zhao (2024) demonstrate larger models exhibit lower perturbation sensitivity. We acknowledge threshold derivation and evidence partially overlap via Lunardi et al.; the thresholds require independent validation on held-out benchmarks (ARC-AGI, SWE-bench) before deployment.

Validation protocol. Compute RR for 20+ models on 500 MMLU questions with 5 paraphrase variants each. Correlate with MMLU → MMLU-Pro degradation. Author-estimated cost: ∼1.5–2× baseline API budget, ∼40 hours researcher time. We have not conducted this validation.

Note on model selection. The empirical demonstration evaluated Claude Sonnet 4, Gemini 2.5 Flash, and GPT-4o—the

three leading frontier model families available via public API at evaluation time. These were not selected post-hoc to maximize the paraphrase sensitivity gap; they represent the full set of models evaluated. The $\sim 40\times$ difference in $\sigma^2_{\text{paraphrase}}$ between two 100%-scoring models was therefore not cherry-picked.

The validation protocol above specifies how to extend this analysis to o-series, Llama, Mistral, and additional families; the community is invited to run it. A wider sweep is needed before treating the sensitivity gap as a general finding rather than an illustrative case.

Goodhart resistance. Static paraphrase sets will eventually be gamed. A natural concern: if paraphrasing becomes standard, models may simply be trained on paraphrased data (data augmentation), treating variance (the symptom) rather than understanding (the disease). We argue this objection is partially addressed by three mechanisms:

1. **Infinite paraphrase space**: Data augmentation covers finite paraphrase variants; evaluation can sample from the unbounded space of semantically-equivalent reformulations. Models trained on paraphrases of MMLU questions may still fail on novel phrasings not seen during augmentation.
2. **Variance pattern signatures**: RR measures the *ratio* of paraphrase-to-seed variance. Data augmentation that improves robustness genuinely (via better representations) should reduce both variances proportionally; augmentation that overfits to known paraphrases may reduce $\sigma^2_{\text{paraphrase}}$ on seen variants while leaving the model fragile to unseen reformulations—detectable via held-out paraphrase sets.
3. **Dynamic generation**: Quality-Controlled Paraphrase Generation (QCPG; Bandel et al., 2022) enables on-the-fly generation with specified semantic similarity and lexical diversity targets. Using a separate paraphrase model (not the model under evaluation) with quality filtering via BERTScore (Zhang et al., 2020) > 0.85 (threshold proposed) provides standardized, non-memorizable variants.

Concrete paraphrase generation protocol. Table 22 specifies the implementation steps for computing RR.

Table 22. Paraphrase generation protocol for computing RR.

Step	Action	Method	Quality Criteria
1	Generate	5 paraphrases per question via QCPG or prompted LLM	Prompt: "Rewrite preserving meaning"
2	Filter	Quality control on generated paraphrases	BERTScore > 0.85; Jaccard < 0.7
3	Evaluate	Model on original + paraphrases	3 random seeds each
4	Compute	$RR = \text{Var(paraphrase)}/\text{Var(seed)}$	—

Author-estimated cost: $\sim 6\times$ API calls per question (1 original + 5 paraphrases), plus one-time paraphrase generation ($\sim$\$50–100 for 500 questions via API).

Defense-in-depth. Validity protocols will become optimization targets. Our tiered approach provides defense-in-depth: gaming RR requires genuine paraphrase robustness (not just memorization of known variants); gaming transfer tests requires actual transfer capability. No single metric is Goodhart-proof, but the combination raises the cost of gaming substantially—a model must simultaneously achieve low variance on dynamic paraphrases *and* transfer to held-out distributions.

O. Distribution Match Protocol

Practitioners need concrete methods to determine whether deployment distributions fall within benchmark scope. We recommend the following protocol, drawing on distribution shift detection literature (Rabanser et al., 2019; Koch et al., 2024).

Interpretation. If all signals are positive, benchmark scores are scope-appropriate. If signals are mixed or negative, transfer tests (Appendix C) are required before capability claims are warranted. These are heuristics, not guarantees; practitioners should use multiple signals.

Table 23. Distribution match protocol: tests and thresholds (author-proposed guidelines requiring validation).

Test	**Method**	**Threshold**	**Interpretation**
Embedding distance	Cosine similarity between deployment queries and benchmark questions (general-purpose encoder)	> 0.7	Distributions likely comparable
Vocabulary overlap	Jaccard similarity of domain-specific terms	> 0.7	Distribution match
Two-sample test	Kolmogorov-Smirnov or Maximum Mean Discrepancy (MMD) on embedding representations	$p > 0.05$	Fail to reject null of identical distributions
Pilot evaluation	Model performance on 100–500 deployment-representative examples	Within 10%	Benchmark scope appropriate